A Kernel Nonconformity Score for Multivariate Conformal Prediction

Louis Meyer
University of Warwick
louis.meyer@warwick.ac.uk

Wenkai Xu
University of Warwick
wenkai.xu@warwick.ac.uk

Abstract

Multivariate conformal prediction requires nonconformity scores that compress residual vectors into scalars while preserving certain implicit geometric structure of the residual distribution. We introduce a Multivariate Kernel Score (MKS) that produces prediction regions that explicitly adapt to this geometry. We show that the proposed score resembles the Gaussian process posterior variance, unifying Bayesian uncertainty quantification with the coverage guarantees of frequentist-type. Moreover, the MKS can be decomposed into an anisotropic Maximum Mean Discrepancy (MMD) that interpolates between kernel density estimation and covariance-weighted distance. We prove finite-sample coverage guarantees and establish convergence rates that depend on the effective rank of the kernel-based covariance operator rather than the ambient dimension, enabling dimension-free adaptation. On regression tasks, the MKS reduces the volume of prediction regions significantly, compared to ellipsoidal baselines while maintaining nominal coverage, with larger gains at higher dimensions and tighter coverage levels.

1 Introduction

Conformal prediction constructs prediction regions with finite-sample coverage guarantees under minimal assumptions on the data distribution [32, 39]. Given a point predictor $\hat{f}$ and a confidence level $1-\alpha$, the framework produces a region $C(X)$ satisfying $\mathbb{P}(Y \in C(X)) \geq 1-\alpha$ for any model class and any distribution [2]. The univariate case is well-understood, conformalized quantile regression [28] and standardized residual methods [19] provide simple and effective prediction intervals. Exact conditional coverage $\mathbb{P}(Y \in C(X) \mid X) \geq 1-\alpha$ cannot be achieved distribution-free without structural assumptions [11, 38]. The efficiency of the score, measured by the size of its prediction region, depends on the choice of nonconformity score and the order it induces.

For multivariate responses $Y \in \mathbb{R}^d$, the nonconformity score determines the underlying "shape" of the prediction region, where the efficiency of the region depends on how faithfully this shape reflects the geometry of the residual distributions. Designing scores that yield data-adaptive regions, without requiring/learning a generative model or imposing a fixed parametric shape, remains an open problem. The difficulty is intrinsic to the multivariate setting. In one dimension, residuals reside in an ordered set $\mathbb{R}$, and admit a natural total order where the quantile is well-defined. In $\mathbb{R}^d$ where no canonical ordering exists, the score function $S : \mathbb{R}^d \to \mathbb{R}$ imposes a total order by compressing a d-dimensional residual into a scalar, and this compression unavoidably discards information. Different choices of S discard different structural properties of the residual distribution, and the efficiency of the prediction region depends predominantly on the information preserved.

Existing approaches split into two families: ellipsoidal approaches and density-based methods. Ellipsoidal approaches score residuals by their Mahalanobis distance under the sample covariance, producing regions aligned with the principal axes of the residual distribution [16, 40]. While still being efficient when the residuals are approximately Gaussian, the ellipsoid must expand into low-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

density areas to maintain coverage when the residual geometry departs from ellipticity, whether from model misspecification, multimodality, or nonlinear dependencies between output coordinates, inflating the region unnecessarily. Density-based methods instead threshold an estimated density, of the residuals or of the response given the covariates, producing level sets that can take arbitrary shape [9, 15]. In practice, nonparametric density estimation deteriorates as the output dimension grows, and even a perfect density estimate ranks residuals by a single scalar that is invariant to directions: two residuals at equal density but in structurally different orientations relative to the calibration data receive identical scores. Ellipsoidal scores capture directional covariance but are restricted to a fixed geometry, while density-based scores capture inherent geometry without preserving the directional information.

We introduce a kernel-based nonconformity score that combines desirable properties from both. The score embeds calibration residuals into a reproducing kernel Hilbert space (RKHS) and measures the conformity of a test residual through a regularized quadratic form involving the empirical covariance operator [12] via *its feature space*. We prove that the proposed score resembles the posterior variance of a Gaussian process (GP) [26] evaluated at the calibration residual locations. Importantly, the GP model does not need to be correctly specified. The posterior variance serves as a scoring function whose geometric properties we inherit from Bayesian nonparametrics; the coverage guarantee rests on conformal calibration, which requires only exchangeability of the scores. Our approach resembles Bayesian geometry for the design of the prediction region while achieving frequentist-type validity.

The MKS identity yields a decomposition into two interpretable components from the kernel literature. The first is a Maximum Mean Discrepancy (MMD) term [13] that measures the discrepancy between the empirical test residual and the calibration distribution, acting as a one-sample typicality test through a kernel density score. The second is a kernel principal component analysis (KPCA) [31] correction term that accounts for the anisotropy of the residual distribution. The prediction regions that result are neither density level sets nor ellipsoids, but kernel-adapted regions shaped by both the density and the directional covariance of the residuals. When the kernel is linear, the score reduces to the regularized Mahalanobis distance, so the framework, in principle, generalizes the ellipsoidal methods of Johnstone et al. [16] and Xu et al. [40].

Our contributions are threefold. (1) We propose MKS: a kernel-based nonconformity score for multivariate conformal prediction and prove that it equals a GP posterior variance (Proposition 4.1). We show that the score decomposes into an MMD term and a KPCA correction (Proposition 4.2); and that the linear kernel recovers the Mahalanobis distance (Proposition 4.4), establishing that our method strictly generalizes ellipsoidal conformal prediction. (2) We provide finite-sample coverage guarantees under exchangeability, with convergence rates depending on the effective rank of the RKHS covariance operator rather than the ambient dimension d (Theorem 3.4); we extend these guarantees to stationary α-mixing sequences (Theorem B.20 in Appendix B.3), following the framework of Xu et al. [40]. (3) We demonstrate on synthetic and real data that the proposed kernel-based score produces smaller prediction regions than Mahalanobis, density-only and optimal transport baselines while maintaining nominal coverage, with the largest gains at higher coverage levels and higher output dimensions. The code of our MKS is available at `https://github.com/meyerlouis/mks-conformal`.

2 Background

Split conformal prediction. We consider a regression setting with inputs $X \in \mathcal{X}$ and multivariate responses $Y \in \mathbb{R}^d$. Given a dataset $\{(X_i, Y_i)\}_{i=1}^n$ drawn from a distribution $\mathbb{P}_{X,Y}$, the goal is to construct a prediction region $C(X_{\text{new}})$ satisfying

$$\mathbb{P}\big(Y_{\text{new}} \in C(X_{\text{new}})\big) \geq 1 - \alpha \tag{1}$$

for a user-specified level $\alpha \in (0, 1)$. Following the split conformal framework [19, 22], the data is randomly partitioned into a training set $\mathcal{D}_{\text{train}}$ of size n_{train} and a calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^T$ of size T. A model $\hat{f} : \mathcal{X} \to \mathbb{R}^d$ is fitted on $\mathcal{D}_{\text{train}}$, and the residuals $\varepsilon_i = Y_i - \hat{f}(X_i) \in \mathbb{R}^d$ are computed on the calibration set $\mathcal{D}_{\text{cal}}$. A nonconformity score function $S : \mathbb{R}^d \to \mathbb{R}$ maps each residual to a scalar measuring its degree of nonconformity, with larger values indicating less typical residuals. The conformal threshold $\hat{q}$ is defined as the $\lceil (1-\alpha)(T+1) \rceil / T$ quantile of the calibration scores $\{S(\varepsilon_1), \ldots, S(\varepsilon_T)\}$, yielding the prediction region

$$C(X) = \big\{ y \in \mathbb{R}^d : S\big(y - \hat{f}(X)\big) \leq \hat{q} \big\} \tag{2}$$

If the calibration scores and the test score are exchangeable, then $C(X)$ satisfies (1) [39]. This guarantee is distribution-free, i.e. the result holds for any data distribution, any model $\hat{f}$, and any score function S. The choice of S does not affect coverage but determines the shape and efficiency of the prediction region, e.g. the length of the prediction interval in one dimension.

Exchangeability. The coverage guarantee (1) requires the calibration residuals $\varepsilon_1, \ldots, \varepsilon_T$ and the test residual ε_{T+1} to be exchangeable. Under the split conformal setup with i.i.d. data, this holds by construction. When the data exhibits temporal or distributional dependence, exchangeability generally fails. Barber et al. [3] provide a general framework for conformal prediction beyond exchangeability, bounding the coverage gap in terms of the total variation distance between the true joint distribution and the nearest exchangeable one. Xu et al. [41] derive finite-sample coverage guarantees under α-mixing dependence, showing that coverage deviations vanish as the mixing coefficients decay. Other forms of non-exchangeability, including covariate shift [36], missing data [42], and contamination [8], have also been studied.

Multivariate nonconformity scores. The choice of S determines the geometry of the prediction region (2), since the region is a sublevel set of S. An intuitive approach applies univariate conformal prediction independently to each coordinate at level α/d, producing hyper-rectangular regions that ignore all the dependencies between outputs. Copula-based methods [20, 34] calibrate the marginals jointly but still yield rectangular regions.

The Mahalanobis distance $S(\varepsilon) = (\varepsilon - \bar{\varepsilon})^\top (\widehat{\Sigma} + \lambda I)^{-1} (\varepsilon - \bar{\varepsilon})$ produces ellipsoidal regions aligned with the principal axes of the residual distribution [16]. Xu et al. [40] use this score with a global covariance estimate for multivariate time series, providing coverage guarantees under α-mixing dependence. Braun et al. [5] learn an input-dependent covariance $\Sigma(X)$, producing locally adaptive ellipsoids with improved conditional coverages. Messoudi et al. [21] estimate the local covariance structures using methods of nearest neighbors.

Density-based approaches use an estimated density as the nonconformity score, producing regions that are its level sets [15, 30]. Methods such as HPD-split [15] estimate the conditional density $\hat{p}(y \mid x)$, so the shape of the region varies with the covariates and targets conditional coverage when the residual distribution depends on X. In practice, nonparametric density estimation becomes unreliable as the output dimension grows, and scalable alternatives require fitting conditional density estimators or generative models [24, 9]. Optimal transport approaches [35, 17] and adaptive norm-based formulations [4] offer more flexible region shapes at the cost of solving expensive optimization problems or training dedicated networks. Dheur et al. [9] provide a comprehensive comparison of multivariate conformal methods within a unified framework.

These methods also differ in whether the prediction region adapts to the covariates. Locally adaptive methods [5, 15, 21] learn a covariate-dependent structure, such as a local covariance or a conditional density, so that the shape of the region changes with X. This targets conditional coverage when the residual distribution depends on the input, at the cost of learning an additional function of X. Our score belongs to the class of methods that produce a single region translated by the point prediction, like the Mahalanobis ellipsoid [16] and the optimal transport score [17]. Within this class, it adapts to the global geometry of the residual distribution, including multimodality, curvature and anisotropy, without learning any covariate-dependent structure.

3 Multivariate Kernel Score

We introduce a kernel-based nonconformity score that extends the Mahalanobis distance to nonlinear settings. Let $\{\varepsilon_i\}_{i=1}^T \subset \mathbb{R}^d$ be calibration residuals obtained from a split conformal procedure, with empirical mean $\bar{\varepsilon} = \frac{1}{T}\sum_{i=1}^T \varepsilon_i$ and sample covariance $\widehat{\Sigma} = \frac{1}{T-1}\sum_{i=1}^T (\varepsilon_i - \bar{\varepsilon})(\varepsilon_i - \bar{\varepsilon})^\top$.

Definition 3.1 (Regularized Mahalanobis Distance)**.** *The regularized Mahalanobis distance nonconformity score is*

$$\mathrm{MD}(\varepsilon) = (\varepsilon - \bar{\varepsilon})^\top \left(\widehat{\Sigma} + \lambda I_d\right)^{-1} (\varepsilon - \bar{\varepsilon}) \tag{3}$$

where $\lambda > 0$ is a regularization parameter.

Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definite kernel with feature map $\phi : \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space satisfying $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}$.

Algorithm 1 Kernel Conformal Prediction

Require: Calibration residuals $\{\varepsilon_i\}_{i=1}^{T}$, kernel k, regularization parameter γ, level α

1: Split $\{1,\dots,T\}$ into disjoint folds $\mathcal{I}_{\text{fit}}$ and $\mathcal{I}_{\text{cal}}$ of equal size, with $T_{\text{fit}} = |\mathcal{I}_{\text{fit}}|$ and $T_{\text{cal}} = |\mathcal{I}_{\text{cal}}|$

2: Precompute $(\tilde{K} + \gamma I_{T_{\text{fit}}})^{-1}$ with centered Gram matrix $\tilde{K}$ built on the fit fold,

$$\tilde{K}_{ij} = k(\varepsilon_i,\varepsilon_j) - \tfrac{1}{T_{\text{fit}}}\sum_l k(\varepsilon_i,\varepsilon_l) - \tfrac{1}{T_{\text{fit}}}\sum_l k(\varepsilon_l,\varepsilon_j) + \tfrac{1}{T_{\text{fit}}^2}\sum_{l,m} k(\varepsilon_l,\varepsilon_m), \quad i,j,l,m \in \mathcal{I}_{\text{fit}}$$

3: **for** $i \in \mathcal{I}_{\text{cal}}$ **do**

4: $\quad \hat{e}_k(\varepsilon_i) \leftarrow \tilde{k}(\varepsilon_i,\varepsilon_i) - \tilde{\mathbf{k}}_*(\varepsilon_i)^\top (\tilde{K} + \gamma I_{T_{\text{fit}}})^{-1} \tilde{\mathbf{k}}_*(\varepsilon_i)$ (as in *Eq.* (5))

5: **end for**

6: $\hat{q} \leftarrow \lceil (1-\alpha)(T_{\text{cal}}+1) \rceil / T_{\text{cal}}$ quantile of $\{\hat{e}_k(\varepsilon_i)\}_{i\in\mathcal{I}_{\text{cal}}}$

Ensure: Prediction region $C(X) = \{y : \hat{e}_k(y - \hat{f}(X)) \le \hat{q}\}$

Definition 3.2 (Multivariate Kernel Score (MKS))**.** *The Multivariate Kernel Score is the regularized Mahalanobis distance computed in the feature space $\mathcal{H}$:*

$$\hat{e}_k(\varepsilon) = \langle \tilde{\phi}(\varepsilon),\ \big(\widehat{C}_\phi + \lambda I_{\mathcal{H}}\big)^{-1} \tilde{\phi}(\varepsilon) \rangle_{\mathcal{H}} \tag{4}$$

where $\tilde{\phi}(\varepsilon) = \phi(\varepsilon) - \hat{\mu}_\phi$ is the centered feature map, $\hat{\mu}_\phi = \frac{1}{T}\sum_{i=1}^{T}\phi(\varepsilon_i)$ is the kernel mean embedding, and $\widehat{C}_\phi = \frac{1}{T-1}\sum_{i=1}^{T}\tilde{\phi}(\varepsilon_i)\otimes\tilde{\phi}(\varepsilon_i)$ is the sample covariance operator on $\mathcal{H}$ [12].

The Tikhonov regularization ensures the score is well-defined. In an RKHS with infinite-dimensional feature space, the covariance operator has eigenvalues decaying to 0 and C_ϕ^{-1} is unbounded, but $(C_\phi + \lambda I_{\mathcal{H}})^{-1}$ has operator norm at most $1/\lambda$. In the Gaussian Process interpretation developed in Section 4, λ corresponds to the variance of the observation noise, controlling the balance between interpolating the calibration data and smoothing over it. Using the kernel trick and the Woodbury matrix identity (proof in Appendix A.1), the score can be computed entirely in terms of kernel evaluations. The procedure is summarized in Algorithm 1.

The score requires a single factorization of $\tilde{K} + \gamma I$ on the fit fold, performed once at $O(T_{\text{fit}}^3)$ time and $O(T_{\text{fit}}^2)$ memory. Each subsequent evaluation then costs $O(T_{\text{fit}}^2)$, and the output dimension only enters through the $O(d)$ cost of each kernel evaluation. For larger calibration sets, the Nyström method approximates $\tilde{K}$ from a subset of the fit fold. Sampling points according to their ridge leverage scores allows the required sample size to scale with the effective dimension of $\tilde{K}$, rather than T_{fit} [1]. By Proposition 4.6, the ridge leverage score of a fit fold point is its score divided by γ, so the points that matter most for the approximation are those the score ranks as most atypical.

Proposition 3.3 (Computable Form)**.** *Let $\tilde{K} \in \mathbb{R}^{T\times T}$ be the centered Gram matrix with entries $\tilde{K}_{ij} = \langle\tilde{\phi}(\varepsilon_i),\tilde{\phi}(\varepsilon_j)\rangle_{\mathcal{H}}$, and let $\tilde{\mathbf{k}}_*(\varepsilon) \in \mathbb{R}^T$ be the centered kernel vector with $[\tilde{\mathbf{k}}_*(\varepsilon)]_i = \langle\tilde{\phi}(\varepsilon),\tilde{\phi}(\varepsilon_i)\rangle_{\mathcal{H}}$. Setting $\gamma = \lambda(T-1)$, the Multivariate Kernel Score is*

$$\hat{e}_k(\varepsilon) = \tilde{k}(\varepsilon,\varepsilon) - \tilde{\mathbf{k}}_*(\varepsilon)^\top (\tilde{K} + \gamma I_T)^{-1} \tilde{\mathbf{k}}_*(\varepsilon) \tag{5}$$

where $\tilde{k}(\varepsilon,\varepsilon) = \|\tilde{\phi}(\varepsilon)\|_{\mathcal{H}}^2$.

Given a target miscoverage level $\alpha \in (0,1)$, we define the conformal threshold $\hat{q}$ as the $\lceil (1-\alpha)(T_{\text{cal}}+1)\rceil/T_{\text{cal}}$ quantile of the scores on the quantile fold, $\{\hat{e}_k(\varepsilon_i)\}_{i\in\mathcal{I}_{\text{cal}}}$. The prediction region for a new input X_{T+1} with point prediction $\hat{f}(X_{T+1})$ is then

$$C(X_{T+1}) = \big\{y \in \mathbb{R}^d : \hat{e}_k\big(y - \hat{f}(X_{T+1})\big) \le \hat{q}\big\} \tag{6}$$

Theorem 3.4 (Coverage Guarantee)**.** *Let the calibration residuals be divided into disjoint folds $\mathcal{I}_{\text{fit}}$ and $\mathcal{I}_{\text{cal}}$, with the score constructed from $\mathcal{I}_{\text{fit}}$ and the threshold taken as the $\lceil(1-\alpha)(|\mathcal{I}_{\text{cal}}|+1)\rceil$-th smallest score on $\mathcal{I}_{\text{cal}}$. If the residuals $\{\varepsilon_i\}_{i\in\mathcal{I}_{\text{cal}}}\cup\{\varepsilon_{T+1}\}$ are exchangeable conditionally on $\mathcal{I}_{\text{fit}}$, then the prediction region* (6) *satisfies*

$$\mathbb{P}\big(Y_{T+1} \in C(X_{T+1})\big) \ge 1-\alpha \tag{7}$$

exactly, with no distributional assumptions on the residual law. Conditionally on $\mathcal{I}_{\mathrm{fit}}$ the score is a fixed measurable function, so exchangeability of the residuals transfers to the scores and (7) *follows from the standard quantile argument.*

Under Assumptions B.1–B.4, which include the standard exogeneity condition $\varepsilon_{T+1} \perp X_{T+1}$, the conditional coverage deviation satisfies

$$\left|\mathbb{P}\big(Y_{T+1} \in C(X_{T+1}) \mid X_{T+1}\big) - (1-\alpha)\right| \leq C \cdot T^{-r} \tag{8}$$

with high probability, where $r > 0$ depends on the estimation rate ρ and the regularization, and C depends on the effective rank $\mathbf{r}(\Sigma) = \mathrm{tr}(\Sigma)/\|\Sigma\|_{\mathrm{op}}$ of the RKHS covariance operator. Under α-mixing we establish marginal coverage guarantees in Appendix B.3.

Remark. For smooth kernels such as the RBF, the eigenvalues of Σ decay rapidly and $\mathbf{r}(\Sigma)$ can be much smaller than both d and T. For any stationary kernel with $k(0) = 1$, the part of the constant C that depends on the kernel does not depend on d (see the remark following Theorem B.13). The detailed convergence analysis of Theorem 3.4 is in Appendix B.

4 Probabilistic and Geometric Structure

4.1 Gaussian Process Connection

The Multivariate Kernel Score has a direct probabilistic interpretation as the posterior variance of a Gaussian process observed at the calibration residuals.

Proposition 4.1 (Gaussian Process Identity)**.** *Let $g \sim \mathcal{GP}(0, \tilde{k})$ be a Gaussian process with zero mean and centered kernel $\tilde{k}$. Given observations at locations $\{\varepsilon_i\}_{i=1}^T$ with noise variance $\sigma_n^2 = \gamma$, the posterior variance at a test point ε is [26, Eq. 2.26]:*

$$\mathrm{Var}_{\mathrm{GP}}\big[g(\varepsilon) \mid \{\varepsilon_i\}_{i=1}^T\big] = \tilde{k}(\varepsilon,\varepsilon) - \tilde{\mathbf{k}}_*(\varepsilon)^\top (\tilde{K} + \gamma I_T)^{-1} \tilde{\mathbf{k}}_*(\varepsilon) \tag{9}$$

Comparing with Proposition 3.3, we have the exact identity

$$\hat{e}_k(\varepsilon) = \mathrm{Var}_{\mathrm{GP}}\big[g(\varepsilon) \mid \{\varepsilon_i\}_{i=1}^T\big] \tag{10}$$

The process g is only introduced for interpretation. It is unrelated to the regression function f and is never observed or estimated. In standard GP regression, the posterior variance quantifies uncertainty about a function at unobserved locations, and depends on the observed function values $\mathbf{z}$ through the posterior mean. Our score depends only on the locations $\{\varepsilon_i\}_{i=1}^T$ of the calibration residuals, not on any response observed at those locations. The MKS is a purely geometric quantity that measures how well a test residual is explained by the spatial configuration of the calibration data. A second distinction is that we use the centered kernel $\tilde{k}$ rather than the raw kernel k, introducing density sensitivity. For the RBF kernel, $k(\varepsilon,\varepsilon) = 1$ for any ε, while after centering, $\tilde{k}(\varepsilon,\varepsilon) = 1 + \bar{k} - 2\hat{p}_k(\varepsilon)$ varies with the local density of the calibration distribution. Centering enables the score to distinguish residuals in dense regions from those in the sparse ones.

The regularization parameter γ enters the GP interpretation as the observation noise variance, and governs how much of the calibration data's structure the score retains. Small γ corresponds to near-interpolation, where the GP posterior variance drops to nearly zero at each calibration point, and the score surface closely tracks the placement of individual residuals. Large γ smooths over individual observations, making the posterior variance reflect only the broad distributional structure, and the score approaches the MMD, as shown in Proposition 4.2.

This interpretation connects Bayesian and frequentist reasoning within a single framework. The GP posterior variance provides geometrically rich machinery for designing the nonconformity score. It captures density, directional covariance, and manifold structure, producing prediction regions that adapt to the shape of the residual distribution. The conformal calibration provides validity, the coverage guarantee (7) holds regardless of whether the GP model is correctly specified, because it relies on exchangeability of the scores, not on Gaussianity of the process.

4.2 MMD-KPCA Decomposition

The Multivariate Kernel Score can also be written in terms of the kernel principal components of the calibration residuals, which separates its density and directional parts.

Proposition 4.2 (MMD–KPCA Decomposition)**.** *Let $Q_T = \frac{1}{T}\sum_{i=1}^{T}\delta_{\varepsilon_i}$ denote the empirical distribution of the calibration residuals, where δ_{ε_i} is the Dirac measure at ε_i (the probability distribution that assigns all its mass to the point ε_i). The Multivariate Kernel Score satisfies*

$$\hat{e}_k(\varepsilon) = \mathrm{MMD}^2(\delta_\varepsilon,\, Q_T) \;-\; \sum_{j:\,\lambda_j>0} \frac{\lambda_j}{\lambda_j+\gamma}\,\beta_j(\varepsilon)^2, \tag{11}$$

where $\mathrm{MMD}^2(\delta_\varepsilon, Q_T) = \|\tilde{\phi}(\varepsilon)\|^2_{\mathcal{H}} = k(\varepsilon,\varepsilon) + \bar{k} - 2\hat{p}_k(\varepsilon)$ is the distance between the Dirac measure at ε and the (empirical) calibration distribution measure in the form of Maximum Mean Discrepancy, with $\hat{p}_k(\varepsilon) = \frac{1}{T}\sum_{i=1}^{T} k(\varepsilon,\varepsilon_i)$ and $\bar{k} = \frac{1}{T^2}\sum_{i,j} k(\varepsilon_i,\varepsilon_j)$. The quantities $\{\lambda_j\}_{j=1}^{T}$ are the eigenvalues of $\tilde{K}$, with corresponding eigenvectors $\{\mathbf{v}_j\}_{j=1}^{T}$, and for each of the $\lambda_j > 0$, $\beta_j(\varepsilon) = \lambda_j^{-1/2}\,\mathbf{v}_j^\top \tilde{\mathbf{k}}_*(\varepsilon)$ is the projection of $\tilde{\phi}(\varepsilon)$ onto the j-th normalized kernel principal component of the calibration data.

The proof is in Appendix A.3. The first term builds on MMD [13], an integral probability metric that measures the distance between two distributions P and Q through their kernel mean embeddings in RKHS, defined as $\mathrm{MMD}^2(P,Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{H}}$. When the kernel is characteristic, $\mathrm{MMD}(P,Q) = 0$ if and only if $P = Q$ [33]. The MMD tests for distributional equality and is widely used as a two-sample test statistic. In our setting, we apply the same principle with $P = \delta_\varepsilon$, a single test residual, and $Q = Q_T$, the empirical calibration distribution, so that the term measures the discrepancy between one point and the calibration set, effectively acting as a one-sample typicality test. For the RBF kernel, $k(\varepsilon,\varepsilon) = 1$ for all ε and $\bar{k}$ depends only on the calibration set, so the only term that varies with the test point is $\hat{p}_k(\varepsilon) = \frac{1}{T}\sum_i k(\varepsilon,\varepsilon_i)$, a Parzen-Rosenblatt kernel density estimate up to normalization [23, 29]. The MMD term is monotonically decreasing in estimated density, implying that the residuals in high-density regions receive low scores and those in sparse regions receive high scores. Using this term alone as a nonconformity score would produce prediction regions that are level sets of the kernel density estimate.

The KPCA correction introduces anisotropy. It reduces the score for residuals whose feature-space representation aligns with directions of high calibration variance, where spread is expected and deviations are uninformative. The score is large only where the calibration data provides neither density support nor directional explanation, and small where either is present. The prediction region is therefore neither a density level set nor an ellipsoid, but shaped by both the local density and the directional covariance of the calibration residuals.

This decomposition mirrors a familiar progression in $\mathbb{R}^d$. The squared error $\|\varepsilon - \bar{\varepsilon}\|^2$ treats all directions equally, and the Mahalanobis distance adds inverse-covariance weighting to penalize deviations along low-variance directions. The MMD term is the RKHS analogue of the squared error, an isotropic distance from the mean embedding that ignores directional structure. The KPCA correction adds inverse-covariance weighting in $\mathcal{H}$, completing the lift from Mahalanobis in $\mathbb{R}^d$ to an anisotropic distance in the feature space. The parameter γ controls how much of the directional structure the score uses. Large γ suppresses the correction and recovers isotropic density-only scoring, while small γ allows the anisotropic structure to shape the prediction region.

4.3 Special Cases and Properties

The two decompositions above yield several properties of the MKS. The score is non-negative, recovers the Mahalanobis distance under a linear kernel, and reduces to ridge leverage scores on the fit fold residuals.

Proposition 4.3 (Non-negativity)**.** *The MKS satisfies $\hat{e}_k(\varepsilon) \geq 0$ for all $\varepsilon \in \mathbb{R}^d$, with equality if and only if $\tilde{\phi}(\varepsilon) = 0$, i.e. $\phi(\varepsilon) = \hat{\mu}_\phi$.*

Proof. Immediate from the GP identity (Proposition 4.1): the score equals the posterior variance of a Gaussian process, which is non-negative by construction. □

Proposition 4.4 (Linear Kernel Recovery)**.** *For the linear kernel $k(\mathbf{x},\mathbf{x}') = \mathbf{x}^\top\mathbf{x}'$, the Multivariate Kernel Score equals $\lambda \cdot \mathrm{MD}(\varepsilon)$, where $\mathrm{MD}(\varepsilon)$ is the regularized Mahalanobis distance of Definition 3.1. Since the constant factor λ does not affect ranking, the two scores produce identical conformal prediction regions.*

The proof is in Appendix A.4. MKS strictly generalizes the Mahalanobis approach [16, 40], as the linear kernel recovers their score exactly, while nonlinear kernels extend the method beyond ellipsoidal regions. In the linear case, the regularization parameter λ corresponds to ridge regularization of the sample covariance, and the calibration scores reduce to the diagonal of matrix $X(X^\top X + \lambda I)^{-1}X^\top$.

Moreover, the Multivariate Kernel Score admits a decomposition into a component within the span of the calibration data in $\mathcal{H}$ and a component orthogonal to it.

Proposition 4.5 (Spectral Decomposition)**.** *Let $\{\lambda_j, \mathbf{v}_j\}_{j=1}^T$ be the eigenvalues and eigenvectors of $\tilde{K}$, and for each $\lambda_j > 0$ let $\mathbf{e}_j = \mathbf{w}_j/\|\mathbf{w}_j\|_{\mathcal{H}} = \mathbf{w}_j/\sqrt{\lambda_j}$ be the normalized kernel principal components in $\mathcal{H}$, where $\mathbf{w}_j = \sum_{i=1}^T [\mathbf{v}_j]_i \tilde{\phi}(\varepsilon_i)$, so that $\beta_j(\varepsilon) = \langle \tilde{\phi}(\varepsilon), \mathbf{e}_j \rangle_{\mathcal{H}}$ with β_j as in Proposition 4.2. Define the orthogonal residual $\tilde{\phi}_\perp(\varepsilon) = \tilde{\phi}(\varepsilon) - \sum_{j:\,\lambda_j>0} \beta_j(\varepsilon)\,\mathbf{e}_j$. Then*

$$\hat{e}_k(\varepsilon) = \sum_{j:\,\lambda_j>0} \frac{\gamma}{\lambda_j + \gamma}\,\beta_j(\varepsilon)^2 + \|\tilde{\phi}_\perp(\varepsilon)\|_{\mathcal{H}}^2 \tag{12}$$

The proof is in Appendix A.5. The within-span component $\sum_j \gamma\,\beta_j(\varepsilon)^2/(\lambda_j+\gamma)$ measures atypicality along the principal directions of the calibration data. Each squared projection is weighted by $\gamma/(\lambda_j + \gamma)$, close to γ/λ_j along directions of large variance, consistent with the Mahalanobis interpretation. The orthogonal component $\|\tilde{\phi}_\perp(\varepsilon)\|_{\mathcal{H}}^2$ captures the part of the test residual that lies entirely outside the span of the calibration data in $\mathcal{H}$. It dominates residuals in regions of the output space not explored by the calibration set, where the data provides no structural information and the GP posterior variance reduces to the prior variance.

The parameter γ controls the balance between these two components. When $\gamma \to 0$, the within-span weights vanish and the score trusts the calibration structure completely and penalizes only the orthogonal component. When $\gamma \to \infty$, the weights approach 1, the KPCA correction in (11) vanishes, and the score reduces to $\mathrm{MMD}^2(\delta_\varepsilon, Q_T)$, recovering isotropic density-only scoring. For fit fold residuals, $\tilde{\phi}(\varepsilon_i)$ lies in the span of the fit fold data by construction, so the orthogonal component vanishes and the score admits a closed-form expression in terms of the kernel ridge regression hat matrix.

Proposition 4.6 (Leverage Score Identity)**.** *Let $H = \tilde{K}(\tilde{K} + \gamma I_T)^{-1}$ be the kernel ridge regression hat matrix, with diagonal entries $h_{ii} = H_{ii}$. For each fit fold residual ε_i, the Multivariate Kernel Score satisfies*

$$\hat{e}_k(\varepsilon_i) = \gamma \cdot h_{ii} \tag{13}$$

The proof is in Appendix A.6. The diagonal entries h_{ii} are the γ-ridge leverage scores of the centered Gram matrix [1], which measure the influence of each calibration point on the kernel ridge regression fit. The conformal threshold is a quantile $\hat{q}$ directly determined by the ordering of the leverages $\{h_{ii}\}_{i=1}^T$, where high-leverage calibration residuals occupying unusual positions in feature space receive high nonconformity scores, and low-leverage residuals in well-populated regions receive low scores. For a test point $\varepsilon \notin \{\varepsilon_i\}$, the orthogonal component $\|\tilde{\phi}_\perp(\varepsilon)\|_{\mathcal{H}}^2$ is generally nonzero, and the MKS can be understood as an out-of-sample extension of leverage, penalizing test residuals that lie outside the span of the calibration data in $\mathcal{H}$.

High-dimensional geometry. Once its shape is fixed by the covariance estimate, the Mahalanobis ellipsoid can only grow through the conformal threshold $\hat{q}$, which scales it in all directions at once. At tight coverage levels, $\hat{q}$ is set by the most extreme calibration residuals, and when the residual distribution is not elliptical these extremes occur in particular directions. Covering them requires enlarging the ellipsoid in every direction, and since its volume grows as $\hat{q}^{\,d/2}$ (Appendix C.1), this cost grows exponentially with the dimension. The MMD term of the MKS assigns high scores wherever calibration support is absent, so the region grows only where the data does. In the case of higher dimensions, each additional output brings new dependencies that move the residual distribution further from any ellipsoid. The MMD term alone, however, is limited by concentration of norm [37, Theorem 6.3.2]. When the residual variance is spread over many directions, pairwise distances between residuals concentrate around a common value. With a lengthscale on the scale of these distances, as the median heuristic gives, the density estimate depends mostly on the distance to the center of the residuals and treats all directions alike. The KPCA correction restores the anisotropy that a covariance captures. Along kernel principal directions of large variance, the correction of

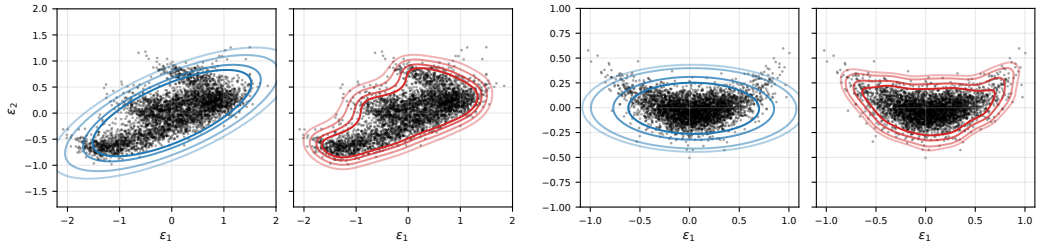
Proposition 4.2 removes almost all of the squared projection, so the score behaves like a Mahalanobis distance in the feature space along these directions, while the MMD term still follows the support of the residuals elsewhere.

5 Experiments

In all experiments, we use the RBF kernel $k(\varepsilon, \varepsilon') = \exp(-\|\varepsilon - \varepsilon'\|^2/2\ell^2)$ for demonstration purposes. The primary contribution of this work is conceptual and theoretical; the choice of kernel and its optimization are important orthogonal questions that we defer to future work. The Multivariate Kernel Score (**MKS**) introduces two hyperparameters, the lengthscale ℓ and the regularization γ. Both are in principle tunable via cross-validation on the calibration scores or through the Gaussian process marginal likelihood implied by Proposition 4.1, and under split calibration both can be selected on the fit fold at no cost to validity. We use simple heuristics throughout the experiments: γ is set to the 90th percentile eigenvalue of the centered Gram matrix $\tilde{K}$ computed on the fit fold, and ℓ is set to half the median pairwise distance between fit fold residuals, the standard heuristic [13]. We compare against the **Mahalanobis score** (Definition 3.1), the **Bonferroni method** (univariate conformal prediction per coordinate at level α/d), the **optimal transport score** of Klein et al. [17], and a **density-only score** using $\mathrm{MMD}^2(\delta_\varepsilon, Q_T)$ alone (the $\gamma \to \infty$ limit of our score). All of these, like our score, produce regions of fixed shape rather than regions that adapt to the covariates. Sensitivity to the lengthscale and the regularization is analyzed in Appendix C.4.4, where coverage holds at the nominal level throughout and the default choice sits in the interior of the region where volumes are stable. Both the kernel score and the density-only score share the same lengthscale, so differences are attributable to the KPCA correction term alone. For each dataset, we fit a **linear model** and a **Multi-Layer Perceptron (MLP)** (Appendix C.3.2) to obtain the residuals and apply all scores to the resulting residuals (see Appendix C.1 for volume estimation details).

5.1 Synthetic Data

We generate $n = 10{,}000$ samples from a bivariate regression task (details in Appendix C.3.1) with non-Gaussian noise exhibiting nonlinear dependence between the output coordinates. We fit both a linear regression and a neural network (2-layer MLP), and apply split conformal prediction with a 50/25/25 train/calibration/test split, the calibration set being further divided into two equal folds, a fit fold on which the score is constructed and a quantile fold on which the threshold is computed. Results are averaged over 30 random seeds.



(a) Linear model residuals

(b) MLP residuals

Figure 1: Prediction regions on synthetic data at coverage levels $1 - \alpha \in \{0.90, 0.95, 0.98, 0.99\}$. In each panel, left: Mahalanobis (blue); right: MKS (red), overlaid on calibration residuals

Figure 1 displays the prediction regions for both models across four coverage levels. Under the linear model (Figure 1a), the residuals exhibit clear non-elliptical structure due to model misspecification. The ellipsoids must expand into low-density regions to achieve the target coverage, wasting volume on regions where residuals are unlikely to occur. The MKS contour follows the geometry of the calibration data, producing tighter regions that exclude these areas. Under the neural network (Figure 1b), the residuals are closer to elliptical but the MKS retains a volume advantage.

Table 1 reports coverage, volume, and worst-slab coverage (WSC) across all methods, models, and coverage levels. The MKS achieves the smallest volume at every α level under both models while maintaining nominal marginal coverage, with WSC reported as an approximation of conditional

α	Method	Linear			NN		
		Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	0.89884±0.01087	4.09042±0.09801	0.70587±0.04177	0.90024±0.01004	0.52885±0.01897	0.86253±0.02117
	Mahal	0.90008±0.01316	2.82873±0.09539	0.77820±0.03143	0.89967±0.01017	0.52723±0.02127	0.86703±0.01798
	OTCP	0.89967±0.01139	2.94837±0.09241	0.77220±0.03784	0.89945±0.01123	0.52438±0.02037	0.85886±0.02320
	Density	0.89749±0.01217	3.15115±0.11313	0.68213±0.05181	0.89937±0.01051	0.56598±0.02350	0.86246±0.02156
	Kernel	0.90085±0.01311	**2.49646±0.08357**	0.81608±0.02734	0.89932±0.01052	**0.47828±0.01605**	0.85339±0.02426
0.05	Bonferroni	0.94961±0.00830	4.97000±0.15966	0.85080±0.03324	0.95141±0.00812	0.72105±0.03409	0.92713±0.01503
	Mahal	0.95021±0.00828	3.56900±0.18571	0.85280±0.02314	0.95020±0.00771	0.75634±0.04378	0.93218±0.01135
	OTCP	0.95024±0.00922	3.60526±0.12250	0.88460±0.02570	0.95028±0.00877	0.72310±0.04076	0.92352±0.01806
	Density	0.95005±0.00916	3.88299±0.14726	0.84880±0.03552	0.95053±0.00896	0.77974±0.04347	0.92633±0.01619
	Kernel	0.95071±0.00714	**3.06383±0.10846**	0.90239±0.01614	0.95119±0.00857	**0.64310±0.03299**	0.91893±0.02269
0.02	Bonferroni	0.98089±0.00484	5.99881±0.24681	0.94186±0.01733	0.98029±0.00604	1.00442±0.08209	0.96793±0.01266
	Mahal	0.98051±0.00450	4.94703±0.35037	0.92787±0.01765	0.98076±0.00481	1.15531±0.10972	0.96973±0.00851
	OTCP	0.97963±0.00483	4.39360±0.14539	0.95193±0.01369	0.97977±0.00410	0.99319±0.05870	0.96453±0.01110
	Density	0.98093±0.00484	4.70010±0.16847	0.94566±0.01715	0.98000±0.00446	1.05032±0.06910	0.96633±0.01088
	Kernel	0.97951±0.00643	**3.83940±0.19987**	0.95633±0.01395	0.97996±0.00535	**0.85463±0.05454**	0.96093±0.01532
0.01	Bonferroni	0.99089±0.00363	6.72434±0.29196	0.97039±0.01352	0.99061±0.00313	1.28914±0.12235	0.98299±0.00538
	Mahal	0.99059±0.00347	6.30128±0.59552	0.96333±0.01394	0.99092±0.00250	1.56025±0.17410	0.98392±0.00366
	OTCP	0.99072±0.00367	5.04567±0.20328	0.97633±0.01096	0.99113±0.00358	1.25061±0.10896	0.98386±0.00682
	Density	0.99060±0.00348	5.29902±0.21313	0.97406±0.01225	0.99129±0.00320	1.29962±0.09400	0.98420±0.00544
	Kernel	0.98987±0.00384	**4.41557±0.24420**	0.97740±0.00964	0.98988±0.00383	**1.04220±0.08911**	0.97893±0.00985

Table 1: Results on synthetic data across 30 seeds. All values reported as mean ± std. Bold indicates best volume per model and α.

coverage[1] (details Appendix C.2). The kernel score produces prediction regions whose volumes are 9 to 33% smaller than Mahalanobis and 15 to 21% smaller than the density-only score. The advantage over the ellipsoid grows at tighter coverage levels. As the target coverage increases, the ellipsoid must extend into the tails of the residual distribution where the calibration data provides little support, and the volume cost of this expansion grows faster for the ellipsoid than for the MKS contour, which expands along the support of the data. This holds under both the misspecified linear model and the neural network, where the distribution is closer to elliptical.

The WSC supports this reading. Under the linear model, the MKS achieves the best WSC at every α level, so the volume reduction reflects a genuinely tighter fit rather than a sacrifice in conditional coverage. The density-only score produces larger regions under both models and at every level, isolating the KPCA correction as the source of the improvement rather than density estimation alone.

5.2 Real Data

We evaluate the method on three regression datasets previously used in Feldman et al. [10] and Romano et al. [28], where we augment the original univariate responses with additional variables from the same source, producing response dimensions $d = 2, 3$, and 4. On the House dataset we further extend the response to $d = 6, 8$ and 10. Volumes are estimated via Monte Carlo (Appendix C.1), and the number of samples required grows exponentially with the output dimension, which limits our evaluation to $d \leq 10$. We leave the evaluation of higher-dimensional responses to future work. All details are provided in Appendix C.4.1. We report results from linear model residuals. MLP regression results are provided in Appendix C.4.3.

Table 2 reports the volume ratio of regions produced using the Multivariate Kernel Score relative to the Mahalanobis baseline, computed over 50 random splits. The MKS produces smaller prediction regions in every dataset, dimension and coverage level, with the largest gains at higher output dimension and tighter coverage. As the number of output coordinates grows, correlations between outputs create residual distributions that an ellipsoid cannot approximate without enclosing large empty regions, and the MKS exploits this structure.

[1]WSC searches over linear slabs, structurally favoring methods whose regions align with linear subspaces, e.g. Mahalanobis ellipsoid and isotropic density level sets. The anisotropic KPCA term can produce boundaries less aligned with linear slabs even when the region is tighter. Comparable WSC at smaller volume therefore indicates a genuinely tighter region rather than a loss of conditional coverage.

Dataset	d	Volume Ratio (MKS / Mahalanobis)				MKS Coverage			
		$1-\alpha$=.90	.95	.98	.99	.90	.95	.98	.99
Bio	2	$0.55604_{\pm 0.141}$	$0.57973_{\pm 0.142}$	$0.62595_{\pm 0.147}$	$0.74917_{\pm 0.166}$	$0.89959_{\pm 0.00639}$	$0.95005_{\pm 0.00416}$	$0.97995_{\pm 0.00234}$	$0.98993_{\pm 0.00167}$
	3	$0.48217_{\pm 0.123}$	$0.47716_{\pm 0.119}$	$0.44091_{\pm 0.103}$	$0.51442_{\pm 0.125}$	$0.89971_{\pm 0.00564}$	$0.95023_{\pm 0.00383}$	$0.97998_{\pm 0.00227}$	$0.99014_{\pm 0.00175}$
	4	$0.40794_{\pm 0.102}$	$0.38616_{\pm 0.096}$	$0.42126_{\pm 0.107}$	$0.34347_{\pm 0.060}$	$0.89978_{\pm 0.00618}$	$0.95023_{\pm 0.00400}$	$0.98017_{\pm 0.00250}$	$0.99004_{\pm 0.00164}$
Blog	2	$0.45632_{\pm 0.051}$	$0.62194_{\pm 0.068}$	$0.75142_{\pm 0.062}$	$0.61701_{\pm 0.075}$	$0.90024_{\pm 0.00457}$	$0.94983_{\pm 0.00352}$	$0.98024_{\pm 0.00216}$	$0.99018_{\pm 0.00152}$
	3	$0.52974_{\pm 0.055}$	$0.58295_{\pm 0.050}$	$0.47078_{\pm 0.053}$	$0.32899_{\pm 0.062}$	$0.89961_{\pm 0.00454}$	$0.95007_{\pm 0.00336}$	$0.98014_{\pm 0.00233}$	$0.99015_{\pm 0.00158}$
	4	$0.23808_{\pm 0.080}$	$0.28461_{\pm 0.086}$	$0.20771_{\pm 0.090}$	$0.11870_{\pm 0.045}$	$0.89926_{\pm 0.00418}$	$0.94993_{\pm 0.00354}$	$0.98017_{\pm 0.00219}$	$0.99001_{\pm 0.00176}$
House	2	$0.87402_{\pm 0.022}$	$0.90569_{\pm 0.028}$	$0.84463_{\pm 0.058}$	$0.70139_{\pm 0.089}$	$0.90038_{\pm 0.00632}$	$0.95031_{\pm 0.00436}$	$0.97987_{\pm 0.00278}$	$0.98999_{\pm 0.00229}$
	3	$0.88351_{\pm 0.027}$	$0.91212_{\pm 0.039}$	$0.74368_{\pm 0.060}$	$0.63522_{\pm 0.076}$	$0.90041_{\pm 0.00700}$	$0.95067_{\pm 0.00448}$	$0.97996_{\pm 0.00277}$	$0.99043_{\pm 0.00202}$
	4	$0.47955_{\pm 0.068}$	$0.51372_{\pm 0.061}$	$0.41236_{\pm 0.076}$	$0.30610_{\pm 0.097}$	$0.90203_{\pm 0.00793}$	$0.95054_{\pm 0.00540}$	$0.98002_{\pm 0.00383}$	$0.99008_{\pm 0.00268}$
	6	$0.22313_{\pm 0.034}$	$0.20704_{\pm 0.032}$	$0.18653_{\pm 0.045}$	$0.14300_{\pm 0.047}$	$0.90138_{\pm 0.00790}$	$0.95037_{\pm 0.00524}$	$0.98023_{\pm 0.00392}$	$0.98999_{\pm 0.00267}$
	8	$0.13958_{\pm 0.035}$	$0.19108_{\pm 0.048}$	$0.11514_{\pm 0.048}$	$0.07368_{\pm 0.037}$	$0.90165_{\pm 0.00743}$	$0.95123_{\pm 0.00551}$	$0.98041_{\pm 0.00387}$	$0.98981_{\pm 0.00268}$
	10	$0.20987_{\pm 0.052}$	$0.29878_{\pm 0.076}$	$0.11135_{\pm 0.053}$	$0.07838_{\pm 0.048}$	$0.90120_{\pm 0.00741}$	$0.95031_{\pm 0.00611}$	$0.98051_{\pm 0.00370}$	$0.98980_{\pm 0.00276}$

Table 2: Real-data results under Ridge residuals, 50 seeds, with split calibration and ℓ=med$/2$. Left: per-seed volume ratio of MKS to Mahalanobis (below 1 = smaller MKS region). Right: MKS empirical coverage (target $1-\alpha$).

Table 2 also confirms that coverage is maintained across all configurations. The empirical coverage of the MKS tracks the nominal level within sampling variability at every dataset, dimension and coverage level. Coverage values are comparable across all methods, consistent with the distribution-free guarantee of the conformal framework. Moreover, all methods achieve comparable approximate conditional coverage in higher dimensions and tighter coverage levels, as measured empirically through the WSC (Appendix C.4.3).

Full results including Bonferroni and density-based nonconformity scores are reported in Appendix C.4.3. The Bonferroni method, ignoring correlations and yielding hyper-rectangles, produces the largest volumes across all settings. Under the median heuristic the density-only score is a strong baseline for $d \leq 4$, improving on Mahalanobis in almost every configuration. In the high-dimensional House settings it falls behind the ellipsoid in most configurations, consistent with the loss of directional information described in Section 4.3. The MKS, which adds the KPCA correction to the same density estimate, produces the smallest regions in almost every configuration. Overall, the Multivariate Kernel Score maintains nominal coverage while reducing region volumes by 9 to 93% relative to the Mahalanobis baseline, with the largest gains at higher dimensions and tighter coverage levels.

6 Conclusion

We introduced the Multivariate Kernel Score (MKS) for multivariate conformal prediction that generalizes the Mahalanobis distance to reproducing kernel Hilbert spaces, producing prediction regions adapted to the residual distribution without imposing a parametric shape. The proposed score can be perceived via two distinct perspectives: (1) the Gaussian process posterior variance; (2) the decomposition into a density term and a directional covariance correction term. The two-sided interpretation of MKS allows us to connect Bayesian uncertainty quantification with distribution-free coverage guarantees of frequentist-type. By compressing multivariate residuals into scalars through the RKHS geometry, the MKS defines a nonparametric center-outward ordering on $\mathbb{R}^d$, with the conformal threshold selecting the quantile level. Finite-sample convergence rates depend on the effective rank of the RKHS covariance operator rather than the ambient dimension. On real data, the MKS significantly reduces prediction region volumes relative to the Mahalanobis baseline while maintaining nominal coverage, with the largest gains at higher dimensions and tighter coverage levels. Future research directions include adaptation to non-exchangeable settings such as incorporating covariate shift, and conformal prediction for function-valued responses.

Acknowledgments and Disclosure of Funding

LM is supported by G-Research through the NextGen Scholarship. WX is partly funded by research grant UKRI3926. We thank Michael Gutmann and Futoshi Futami for helpful discussions.

References

[1] Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel methods with statistical guarantees, 2015.

[2] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022.

[3] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

[4] Sacha Braun, Liviu Aolaritei, Michael I Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression. *arXiv preprint arXiv:2503.19068*, 2025.

[5] Sacha Braun, Eugène Berta, Michael I Jordan, and Francis Bach. Multivariate conformal prediction via conformalized gaussian scoring. *arXiv preprint arXiv:2507.20941*, 2025.

[6] Krisztian Buza. BlogFeedback. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C58S3F.

[7] Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22(81):1–42, 2021.

[8] Jason Clarkson, Wenkai Xu, Mihai Cucuringu, and Gesine Reinert. Split conformal prediction under data contamination. *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, pages 5–27, 2024.

[9] Victor Dheur, Matteo Fontana, Yorick Estievenart, Naomi Desobry, and Souhaib Ben Taieb. A unified comparative study with generalized conformity scores for multi-output conformal regression. *arXiv preprint arXiv:2501.10533*, 2025.

[10] Shai Feldman, Stephen Bates, and Yaniv Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

[11] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

[12] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

[13] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

[14] Harlfoxem. House sales in king county, USA. `https://www.kaggle.com/harlfoxem/housesalesprediction/metadata`. Accessed: July, 2021.

[15] Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022.

[16] Chancellor Johnstone and Bruce Cox. Conformal uncertainty sets for robust optimization. In *Conformal and Probabilistic Prediction and Applications*, pages 72–90. PMLR, 2021.

[17] Michal Klein, Louis Bethune, Eugene Ndiaye, and Marco Cuturi. Multivariate conformal prediction using optimal transport. *arXiv preprint arXiv:2502.03609*, 2025.

[18] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

[19] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[20] Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101, 2021.

[21] Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Ellipsoidal conformal inference for multi-target regression. In *Conformal and Probabilistic Prediction with Applications*, pages 294–306. PMLR, 2022.

[22] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pages 345–356. Springer, 2002.

[23] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[24] Vincent Plassier, Alexander Fishkov, Mohsen Guizani, Maxim Panov, and Eric Moulines. Probabilistic conformal prediction with approximate conditional validity. *arXiv preprint arXiv:2407.01794*, 2024.

[25] Prashant Rana. Physicochemical Properties of Protein Tertiary Structure. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5QW3H.

[26] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[27] Emmanuel Rio. *Asymptotic Theory of Weakly Dependent Random Processes*. Springer, 2017.

[28] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[29] Murray Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[30] Max Sampson and Kung-Sik Chan. Flexible conformal highest predictive conditional density sets, 2025.

[31] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[32] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

[33] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

[34] Sophia Huiwen Sun and Rose Yu. Copula conformal prediction for multi-step time series prediction. In *The Twelfth International Conference on Learning Representations*, 2024.

[35] Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. *arXiv preprint arXiv:2501.18991*, 2025.

[36] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

[37] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[38] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

[39] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[40] Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets, 2024.

[41] Chen Xu and Yao Xie. Conformal prediction for time series. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):11575–11587, 2023.

[42] Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. Conformal prediction with missing values. In *International Conference on Machine Learning*, pages 40578–40604, 2023.

A Derivations and Proofs

A.1 Derivation of Proposition 3.3 (Computable Form)

We show that the RKHS Mahalanobis distance in Definition 3.2 reduces to the finite-dimensional expression in Proposition 3.3 via the Woodbury matrix identity.

Proof. Let $\Phi = [\phi(\varepsilon_1), \dots, \phi(\varepsilon_T)]$ denote the collection of feature vectors in $\mathcal{H}$, and let $H = I_T - \frac{1}{T}\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{T\times T}$ be the centering matrix. The centered feature vectors are $\tilde{\Phi} = \Phi H$, and the sample covariance operator is $\widehat{C}_\phi = \frac{1}{T-1}\tilde{\Phi}\tilde{\Phi}^\top$.

We apply the Woodbury matrix identity. For $A = \lambda I_{\mathcal{H}}$, $U = \frac{1}{\sqrt{T-1}}\tilde{\Phi}$, and $C = I_T$, the identity $(A + UU^\top)^{-1} = A^{-1} - A^{-1}U(I + U^\top A^{-1}U)^{-1}U^\top A^{-1}$ gives:

$$\left(\widehat{C}_\phi + \lambda I_{\mathcal{H}}\right)^{-1} = \frac{1}{\lambda} I_{\mathcal{H}} - \frac{1}{\lambda^2(T-1)}\tilde{\Phi}\left(I_T + \frac{1}{\lambda(T-1)}\tilde{\Phi}^\top\tilde{\Phi}\right)^{-1}\tilde{\Phi}^\top \tag{14}$$

The centered Gram matrix is $\tilde{K} = \tilde{\Phi}^\top\tilde{\Phi} = HKH \in \mathbb{R}^{T\times T}$, where $K_{ij} = k(\varepsilon_i, \varepsilon_j)$. Substituting into (14):

$$\left(\widehat{C}_\phi + \lambda I_{\mathcal{H}}\right)^{-1} = \frac{1}{\lambda} I_{\mathcal{H}} - \frac{1}{\lambda^2(T-1)}\tilde{\Phi}\left(I_T + \frac{1}{\lambda(T-1)}\tilde{K}\right)^{-1}\tilde{\Phi}^\top \tag{15}$$

Let $\tilde{\phi}(\varepsilon) = \phi(\varepsilon) - \hat{\mu}_\phi$ be the centered feature map of a test residual. Applying Definition 3.2:

$$\begin{aligned}\hat{e}_k(\varepsilon) &= \left\langle \tilde{\phi}(\varepsilon), \left(\widehat{C}_\phi + \lambda I_{\mathcal{H}}\right)^{-1}\tilde{\phi}(\varepsilon)\right\rangle_{\mathcal{H}} \\ &= \frac{1}{\lambda}\tilde{k}(\varepsilon,\varepsilon) - \frac{1}{\lambda^2(T-1)}\tilde{\mathbf{k}}_*(\varepsilon)^\top\left(I_T + \frac{1}{\lambda(T-1)}\tilde{K}\right)^{-1}\tilde{\mathbf{k}}_*(\varepsilon)\end{aligned} \tag{16}$$

where $\tilde{\mathbf{k}}_*(\varepsilon) = \tilde{\Phi}^\top\tilde{\phi}(\varepsilon) \in \mathbb{R}^T$ is the centered kernel vector with entries $[\tilde{\mathbf{k}}_*(\varepsilon)]_i = \langle\tilde{\phi}(\varepsilon), \tilde{\phi}(\varepsilon_i)\rangle_{\mathcal{H}}$.

Setting $\gamma = \lambda(T-1)$, we have $I_T + \frac{1}{\gamma}\tilde{K} = \frac{1}{\gamma}(\tilde{K} + \gamma I_T)$, so $(I_T + \frac{1}{\gamma}\tilde{K})^{-1} = \gamma(\tilde{K} + \gamma I_T)^{-1}$. Substituting into (16):

$$\begin{aligned}\hat{e}_k(\varepsilon) &= \frac{1}{\lambda}\tilde{k}(\varepsilon,\varepsilon) - \frac{\gamma}{\lambda^2(T-1)}\tilde{\mathbf{k}}_*(\varepsilon)^\top(\tilde{K} + \gamma I_T)^{-1}\tilde{\mathbf{k}}_*(\varepsilon) \\ &= \frac{1}{\lambda}\left[\tilde{k}(\varepsilon,\varepsilon) - \tilde{\mathbf{k}}_*(\varepsilon)^\top(\tilde{K} + \gamma I_T)^{-1}\tilde{\mathbf{k}}_*(\varepsilon)\right]\end{aligned}$$

where the second line uses $\gamma/[\lambda(T-1)] = 1$. Since the factor $1/\lambda$ is a positive constant that does not affect the ranking of residuals, we define the Multivariate Kernel Score as in Proposition 3.3:

$$\hat{e}_k(\varepsilon) = \tilde{k}(\varepsilon,\varepsilon) - \tilde{\mathbf{k}}_*(\varepsilon)^\top(\tilde{K} + \gamma I_T)^{-1}\tilde{\mathbf{k}}_*(\varepsilon)$$

For completeness, the centered quantities in terms of the raw kernel are:

$$\tilde{k}(\varepsilon,\varepsilon) = k(\varepsilon,\varepsilon) - \frac{2}{T}\sum_{i=1}^{T} k(\varepsilon,\varepsilon_i) + \frac{1}{T^2}\sum_{i,j=1}^{T} k(\varepsilon_i,\varepsilon_j),$$

$$[\tilde{\mathbf{k}}_*(\varepsilon)]_i = k(\varepsilon,\varepsilon_i) - \frac{1}{T}\sum_{j=1}^{T} k(\varepsilon,\varepsilon_j) - \frac{1}{T}\sum_{j=1}^{T} k(\varepsilon_j,\varepsilon_i) + \frac{1}{T^2}\sum_{j,\ell=1}^{T} k(\varepsilon_j,\varepsilon_\ell)$$

□

A.2 Proof of Proposition 4.1 (Gaussian Process Identity)

Proof. Let $g \sim \mathcal{GP}(0, \tilde{k})$ and suppose we observe g at locations $\varepsilon_1, \ldots, \varepsilon_T$ with i.i.d. Gaussian noise of variance γ. By standard GP conditioning [26, Eq. 2.26], the posterior variance at a test point ε is

$$\mathrm{Var}_{\mathrm{GP}}\big[g(\varepsilon) \mid \{\varepsilon_i\}_{i=1}^T\big] = \tilde{k}(\varepsilon, \varepsilon) - \tilde{\mathbf{k}}_*(\varepsilon)^\top (\tilde{K} + \gamma I_T)^{-1} \tilde{\mathbf{k}}_*(\varepsilon)$$

where $\tilde{K}_{ij} = \tilde{k}(\varepsilon_i, \varepsilon_j)$ and $[\tilde{\mathbf{k}}_*(\varepsilon)]_i = \tilde{k}(\varepsilon, \varepsilon_i)$. This is identical to the expression in Proposition 3.3. Note that the posterior variance depends only on the observation locations and the kernel, not on the observed function values. The identity $\hat{e}_k(\varepsilon) = \mathrm{Var}_{\mathrm{GP}}[g(\varepsilon) \mid \{\varepsilon_i\}_{i=1}^T]$ follows by direct comparison. □

A.3 Proof of Proposition 4.2 (MMD–KPCA Decomposition)

Proof. We establish two intermediate steps and combine them for the result.

The MMD term. The squared maximum mean discrepancy between two distributions P and Q with respect to a kernel k is [13]

$$\mathrm{MMD}^2(P, Q) = \mathbb{E}_{x,x' \sim P}\big[k(x, x')\big] + \mathbb{E}_{y,y' \sim Q}\big[k(y, y')\big] - 2\,\mathbb{E}_{x \sim P,\, y \sim Q}\big[k(x, y)\big]$$

For $P = \delta_\varepsilon$ (the Dirac measure at ε) and $Q = Q_T = \frac{1}{T}\sum_{i=1}^T \delta_{\varepsilon_i}$ (the empirical measure of the calibration residuals), each expectation evaluates as follows. Since $x = x' = \varepsilon$ almost surely under δ_ε: $\mathbb{E}_{x,x' \sim \delta_\varepsilon}[k(x, x')] = k(\varepsilon, \varepsilon)$. Since y, y' are drawn independently from Q_T: $\mathbb{E}_{y,y' \sim Q_T}[k(y, y')] = \frac{1}{T^2}\sum_{i,j} k(\varepsilon_i, \varepsilon_j) = \bar{k}$. The cross term: $\mathbb{E}_{x \sim \delta_\varepsilon, y \sim Q_T}[k(x, y)] = \frac{1}{T}\sum_i k(\varepsilon, \varepsilon_i) = \hat{p}_k(\varepsilon)$. Therefore

$$\mathrm{MMD}^2(\delta_\varepsilon, Q_T) = k(\varepsilon, \varepsilon) + \bar{k} - 2\hat{p}_k(\varepsilon)$$

Equivalently, $\mathrm{MMD}^2(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}^2$ where $\mu_P = \mathbb{E}_{x \sim P}[\phi(x)]$ is the kernel mean embedding [13]. Since $\mu_{\delta_\varepsilon} = \phi(\varepsilon)$ and $\mu_{Q_T} = \hat{\mu}_\phi$, we have

$$\mathrm{MMD}^2(\delta_\varepsilon, Q_T) = \|\phi(\varepsilon) - \hat{\mu}_\phi\|_{\mathcal{H}}^2 = \|\tilde{\phi}(\varepsilon)\|_{\mathcal{H}}^2 = \tilde{k}(\varepsilon, \varepsilon)$$

The KPCA correction term. Let $\tilde{K} = V \Lambda V^\top$ be the eigendecomposition, with $V = [\mathbf{v}_1, \ldots, \mathbf{v}_T]$ orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_T)$. Then

$$(\tilde{K} + \gamma I_T)^{-1} = V(\Lambda + \gamma I_T)^{-1} V^\top = \sum_{j=1}^T \frac{1}{\lambda_j + \gamma}\, \mathbf{v}_j \mathbf{v}_j^\top$$

and therefore

$$\tilde{\mathbf{k}}_*(\varepsilon)^\top (\tilde{K} + \gamma I_T)^{-1} \tilde{\mathbf{k}}_*(\varepsilon) = \sum_{j=1}^T \frac{(\mathbf{v}_j^\top \tilde{\mathbf{k}}_*(\varepsilon))^2}{\lambda_j + \gamma} = \sum_{j=1}^T \frac{\alpha_j(\varepsilon)^2}{\lambda_j + \gamma}$$

where $\alpha_j(\varepsilon) = \mathbf{v}_j^\top \tilde{\mathbf{k}}_*(\varepsilon)$.

Verification that $\alpha_j(\varepsilon)$ is a KPCA projection. Define the RKHS vectors $\mathbf{w}_j = \sum_{i=1}^T [\mathbf{v}_j]_i\, \tilde{\phi}(\varepsilon_i) \in \mathcal{H}$. These are eigenfunctions of the sample covariance operator $\widehat{C}_\phi$ with eigenvalues $\lambda_j/(T-1)$:

$$\begin{aligned}
\widehat{C}_\phi\, \mathbf{w}_j &= \frac{1}{T-1} \sum_{\ell=1}^T \tilde{\phi}(\varepsilon_\ell) \left\langle \tilde{\phi}(\varepsilon_\ell),\, \sum_{i=1}^T [\mathbf{v}_j]_i\, \tilde{\phi}(\varepsilon_i) \right\rangle_{\mathcal{H}} \\
&= \frac{1}{T-1} \sum_{\ell=1}^T \tilde{\phi}(\varepsilon_\ell) \sum_{i=1}^T [\mathbf{v}_j]_i\, \tilde{K}_{\ell i} = \frac{1}{T-1} \sum_{\ell=1}^T \tilde{\phi}(\varepsilon_\ell)\, [\tilde{K}\mathbf{v}_j]_\ell \\
&= \frac{\lambda_j}{T-1} \sum_{\ell=1}^T [\mathbf{v}_j]_\ell\, \tilde{\phi}(\varepsilon_\ell) = \frac{\lambda_j}{T-1}\, \mathbf{w}_j
\end{aligned}$$

The projection of $\tilde{\phi}(\varepsilon)$ onto $\mathbf{w}_j$ is

$$\langle \tilde{\phi}(\varepsilon),\, \mathbf{w}_j \rangle_{\mathcal{H}} = \sum_{i=1}^T [\mathbf{v}_j]_i\, \langle \tilde{\phi}(\varepsilon),\, \tilde{\phi}(\varepsilon_i) \rangle_{\mathcal{H}} = \mathbf{v}_j^\top \tilde{\mathbf{k}}_*(\varepsilon) = \alpha_j(\varepsilon)$$

confirming that $\alpha_j(\varepsilon)$ is the projection of $\tilde{\phi}(\varepsilon)$ onto the j-th kernel principal component.

Combining. From Proposition 3.3:

$$\hat{e}_k(\varepsilon) = \tilde{k}(\varepsilon,\varepsilon) - \tilde{\mathbf{k}}_*(\varepsilon)^\top (\tilde{K} + \gamma I_T)^{-1} \tilde{\mathbf{k}}_*(\varepsilon) = \mathrm{MMD}^2(\delta_\varepsilon, Q_T) - \sum_{j=1}^{T} \frac{\alpha_j(\varepsilon)^2}{\lambda_j + \gamma}$$

Here, for $\lambda_j > 0$, $\beta_j(\varepsilon) = \lambda_j^{-1/2}\,\alpha_j(\varepsilon) = \langle \tilde{\phi}(\varepsilon), \mathbf{w}_j/\sqrt{\lambda_j}\rangle_{\mathcal{H}}$ is the projection of $\tilde{\phi}(\varepsilon)$ onto the normalized principal component $\mathbf{w}_j/\sqrt{\lambda_j}$, which has unit norm since $\|\mathbf{w}_j\|_{\mathcal{H}}^2 = \mathbf{v}_j^\top \tilde{K}\,\mathbf{v}_j = \lambda_j$. For $\lambda_j = 0$, $\|\mathbf{w}_j\|_{\mathcal{H}}^2 = \mathbf{v}_j^\top \tilde{K}\,\mathbf{v}_j = \lambda_j = 0$, so $\mathbf{w}_j = 0$ and $\alpha_j(\varepsilon) = \langle \tilde{\phi}(\varepsilon), \mathbf{w}_j\rangle_{\mathcal{H}} = 0$, and these terms vanish. For $\lambda_j > 0$, $\alpha_j(\varepsilon) = \sqrt{\lambda_j}\,\beta_j(\varepsilon)$, hence

$$\frac{\alpha_j(\varepsilon)^2}{\lambda_j + \gamma} = \frac{\lambda_j}{\lambda_j + \gamma}\,\beta_j(\varepsilon)^2.$$

Substituting into the above gives

$$\hat{e}_k(\varepsilon) = \mathrm{MMD}^2(\delta_\varepsilon, Q_T) - \sum_{j:\,\lambda_j > 0} \frac{\lambda_j}{\lambda_j + \gamma}\,\beta_j(\varepsilon)^2. \qquad \square$$

A.4 Proof of Proposition 4.4 (Linear Kernel Recovery)

Proof. For the linear kernel $k(\mathbf{x},\mathbf{x}') = \mathbf{x}^\top\mathbf{x}'$, the feature map is $\phi(\mathbf{x}) = \mathbf{x}$, so $\hat{\mu}_\phi = \bar{\varepsilon}$, $\tilde{\phi}(\varepsilon_i) = \varepsilon_i - \bar{\varepsilon}$, and $\tilde{K}_{ij} = (\varepsilon_i - \bar{\varepsilon})^\top(\varepsilon_j - \bar{\varepsilon})$.

Let $\tilde{E} = [\varepsilon_1 - \bar{\varepsilon}, \ldots, \varepsilon_T - \bar{\varepsilon}] \in \mathbb{R}^{d\times T}$. Then $\tilde{K} = \tilde{E}^\top\tilde{E}$, $\tilde{\mathbf{k}}_*(\varepsilon) = \tilde{E}^\top(\varepsilon - \bar{\varepsilon})$, and $\tilde{k}(\varepsilon,\varepsilon) = \|\varepsilon - \bar{\varepsilon}\|^2$. The score in Proposition 3.3 becomes:

$$\begin{aligned}\hat{e}_k(\varepsilon) &= \|\varepsilon - \bar{\varepsilon}\|^2 - (\varepsilon - \bar{\varepsilon})^\top \tilde{E}\,(\tilde{E}^\top\tilde{E} + \gamma I_T)^{-1}\tilde{E}^\top(\varepsilon - \bar{\varepsilon}) \\ &= (\varepsilon - \bar{\varepsilon})^\top \left[I_d - \tilde{E}\,(\tilde{E}^\top\tilde{E} + \gamma I_T)^{-1}\tilde{E}^\top\right](\varepsilon - \bar{\varepsilon}) \qquad (17)\end{aligned}$$

By the push-through identity, $\tilde{E}\,(\tilde{E}^\top\tilde{E} + \gamma I_T)^{-1}\tilde{E}^\top = (\tilde{E}\,\tilde{E}^\top + \gamma I_d)^{-1}\tilde{E}\,\tilde{E}^\top$. Substituting into (17):

$$\begin{aligned}\hat{e}_k(\varepsilon) &= (\varepsilon - \bar{\varepsilon})^\top \left[I_d - (\tilde{E}\,\tilde{E}^\top + \gamma I_d)^{-1}\tilde{E}\,\tilde{E}^\top\right](\varepsilon - \bar{\varepsilon}) \\ &= (\varepsilon - \bar{\varepsilon})^\top (\tilde{E}\,\tilde{E}^\top + \gamma I_d)^{-1}\left[(\tilde{E}\,\tilde{E}^\top + \gamma I_d) - \tilde{E}\,\tilde{E}^\top\right](\varepsilon - \bar{\varepsilon}) \\ &= \gamma\cdot(\varepsilon - \bar{\varepsilon})^\top (\tilde{E}\,\tilde{E}^\top + \gamma I_d)^{-1}(\varepsilon - \bar{\varepsilon})\end{aligned}$$

Since $\tilde{E}\,\tilde{E}^\top = (T-1)\widehat{\Sigma}$ and $\gamma = \lambda(T-1)$:

$$\begin{aligned}\hat{e}_k(\varepsilon) &= \lambda(T-1)\cdot(\varepsilon - \bar{\varepsilon})^\top\left[(T-1)(\widehat{\Sigma} + \lambda I_d)\right]^{-1}(\varepsilon - \bar{\varepsilon}) \\ &= \lambda\cdot(\varepsilon - \bar{\varepsilon})^\top(\widehat{\Sigma} + \lambda I_d)^{-1}(\varepsilon - \bar{\varepsilon}) = \lambda\cdot\mathrm{MD}(\varepsilon)\end{aligned}$$

The constant factor $\lambda > 0$ does not affect the ranking of residuals, so the MKS and the regularized Mahalanobis distance produce identical conformal prediction regions. $\square$

A.5 Proof of Proposition 4.5 (Spectral Decomposition)

Proof. Throughout, sums run over the indices with $\lambda_j > 0$, since the remaining terms vanish by the proof of Proposition 4.2. From the proof of Proposition 4.2, the projection of $\tilde{\phi}(\varepsilon)$ onto $\mathbf{w}_j$ satisfies $\langle \tilde{\phi}(\varepsilon), \mathbf{w}_j\rangle_{\mathcal{H}} = \alpha_j(\varepsilon)$, so the projection onto the normalized direction is

$$\langle \tilde{\phi}(\varepsilon), \mathbf{e}_j\rangle_{\mathcal{H}} = \frac{\alpha_j(\varepsilon)}{\sqrt{\lambda_j}}$$

Since the $\mathbf{e}_j$ are orthonormal and $\tilde{\phi}_\perp(\varepsilon)$ is orthogonal to all of them, Pythagoras' theorem gives

$$\|\tilde{\phi}(\varepsilon)\|_{\mathcal{H}}^2 = \sum_{j=1}^{T} \langle \tilde{\phi}(\varepsilon), \mathbf{e}_j \rangle_{\mathcal{H}}^2 + \|\tilde{\phi}_\perp(\varepsilon)\|_{\mathcal{H}}^2 = \sum_{j=1}^{T} \frac{\alpha_j(\varepsilon)^2}{\lambda_j} + \|\tilde{\phi}_\perp(\varepsilon)\|_{\mathcal{H}}^2$$

Substituting into the MMD–KPCA decomposition (Proposition 4.2), with $\|\tilde{\phi}(\varepsilon)\|_{\mathcal{H}}^2 = \mathrm{MMD}^2(\delta_\varepsilon, Q_T)$:

$$\begin{aligned}
\hat{e}_k(\varepsilon) &= \|\tilde{\phi}(\varepsilon)\|_{\mathcal{H}}^2 - \sum_{j=1}^{T} \frac{\alpha_j(\varepsilon)^2}{\lambda_j + \gamma} \\
&= \sum_{j=1}^{T} \frac{\alpha_j(\varepsilon)^2}{\lambda_j} + \|\tilde{\phi}_\perp(\varepsilon)\|_{\mathcal{H}}^2 - \sum_{j=1}^{T} \frac{\alpha_j(\varepsilon)^2}{\lambda_j + \gamma} \\
&= \sum_{j=1}^{T} \alpha_j(\varepsilon)^2 \left[\frac{1}{\lambda_j} - \frac{1}{\lambda_j + \gamma} \right] + \|\tilde{\phi}_\perp(\varepsilon)\|_{\mathcal{H}}^2 \\
&= \sum_{j=1}^{T} \frac{\gamma \cdot \alpha_j(\varepsilon)^2}{\lambda_j(\lambda_j + \gamma)} + \|\tilde{\phi}_\perp(\varepsilon)\|_{\mathcal{H}}^2
\end{aligned}$$

Since $\alpha_j(\varepsilon)^2/\lambda_j = \beta_j(\varepsilon)^2$ for $\lambda_j > 0$, this gives

$$\hat{e}_k(\varepsilon) = \sum_{j:\, \lambda_j > 0} \frac{\gamma}{\lambda_j + \gamma} \beta_j(\varepsilon)^2 + \|\tilde{\phi}_\perp(\varepsilon)\|_{\mathcal{H}}^2. \qquad \square$$

A.6 Proof of Proposition 4.6 (Leverage Score Identity)

Proof. For a calibration residual ε_i, $\tilde{\phi}(\varepsilon_i)$ lies in the span of the calibration data in $\mathcal{H}$ by construction, so $\|\tilde{\phi}_\perp(\varepsilon_i)\|_{\mathcal{H}}^2 = 0$ and the spectral decomposition (Proposition 4.5) reduces to

$$\hat{e}_k(\varepsilon_i) = \sum_{j=1}^{T} \frac{\gamma\, \alpha_j(\varepsilon_i)^2}{\lambda_j(\lambda_j + \gamma)}.$$

Since $\tilde{\mathbf{k}}_*(\varepsilon_i) = \tilde{K}_{:,i}$ and $\tilde{K}\mathbf{v}_j = \lambda_j \mathbf{v}_j$, the KPCA projection at a calibration point is $\alpha_j(\varepsilon_i) = \mathbf{v}_j^\top \tilde{K}_{:,i} = \lambda_j [\mathbf{v}_j]_i$. Substituting,

$$\hat{e}_k(\varepsilon_i) = \sum_{j=1}^{T} \frac{\gamma\, \lambda_j\, [\mathbf{v}_j]_i^2}{\lambda_j + \gamma} = \gamma \sum_{j=1}^{T} \frac{\lambda_j}{\lambda_j + \gamma} [\mathbf{v}_j]_i^2 = \gamma \cdot h_{ii}$$

where $h_{ii} = [V\Lambda(\Lambda + \gamma I_T)^{-1} V^\top]_{ii}$ is the i-th diagonal entry of the matrix $H = \tilde{K}(\tilde{K} + \gamma I_T)^{-1}$. $\square$

B Theoretical Analysis

This section provides the full theoretical analysis for the coverage guarantees stated in Theorem 3.4. We first state the assumptions and auxiliary results, then prove asymptotic conditional coverage under i.i.d. residuals (Section B.2), where the standard exogeneity condition makes the conditional guarantee valid, and finally establish marginal coverage under α-mixing dependence (Section B.3), where temporal correlation between X_{T+1} and ε_{T+1} precludes a conditional guarantee.

Throughout this section, we work at the RKHS operator level. The true nonconformity score at time t is

$$e_t = \langle \phi(\varepsilon_t) - \mu,\ \Sigma_\lambda^{-1}(\phi(\varepsilon_t) - \mu) \rangle_{\mathcal{H}}$$

and the estimated nonconformity score is

$$\hat{e}_t = \langle \phi(\hat{\varepsilon}_t) - \hat{\mu},\ \widehat{\Sigma}_\lambda^{-1}(\phi(\hat{\varepsilon}_t) - \hat{\mu}) \rangle_{\mathcal{H}}$$

where $\Sigma_\lambda = \Sigma + \lambda I_{\mathcal{H}}$ and $\widehat{\Sigma}_\lambda = \widehat{\Sigma} + \lambda I_{\mathcal{H}}$ are the regularized true and empirical cross-covariance operators. Here μ and Σ are the population mean embedding and covariance operator (fixed), while $\widehat{\mu}$ and $\widehat{\Sigma}$ are estimated from the calibration set. The estimated score $\hat{e}_t$ corresponds to $\hat{e}_k(\hat{\varepsilon}_t)$ in Proposition 3.3 after the Woodbury reduction (Appendix A.1).

We also define the empirical CDFs of the true and estimated scores:

$$\widetilde{F}_{T+1}(x) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{e_t \le x\}} \qquad \widehat{F}_{T+1}(x) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{\hat{e}_t \le x\}}.$$

B.1 Assumptions

Assumption B.1 (i.i.d. Residuals and Exogeneity)**.** *The residuals $\{\varepsilon_t\}_{t=1}^{T+1}$ are independent and identically distributed. Furthermore, the CDF F_e of the true nonconformity score e_t is Lipschitz continuous with constant $L > 0$. Finally, the test residual ε_{T+1} is independent of the test input X_{T+1}.*

Remark. The Lipschitz condition on F_e ensures that small perturbations of the scores lead to small changes in coverage probability. This assumption is relaxed in Section B.3, where we extend the results to stationary α-mixing sequences. The exogeneity condition $\varepsilon_{T+1} \perp X_{T+1}$ is standard in regression and ensures that the true nonconformity score e_{T+1} is independent of X_{T+1}, which allows the marginal bound in the proof to hold conditionally on any fixed $X_{T+1} = x$. Under α-mixing dependence (Section B.3), X_{T+1} typically contains lags of Y_t and is correlated with ε_{T+1}, so this condition fails and only marginal coverage is established there.

Assumption B.2 (Estimation Quality in RKHS)**.** *There exists a sequence $\{\delta_T\}_{T\ge 1}$ satisfying $\delta_T = O(T^{-\rho})$ for some $\rho > 0$ such that:*

$$\frac{1}{T}\sum_{t=1}^{T} \|\phi(\hat{\varepsilon}_t) - \phi(\varepsilon_t)\|_{\mathcal{H}}^2 \le \delta_T^2 \qquad \textit{and} \qquad \|\phi(\hat{\varepsilon}_{T+1}) - \phi(\varepsilon_{T+1})\|_{\mathcal{H}} \le \delta_T.$$

Remark. This requires the prediction error in the RKHS feature space to be small on average over the calibration set and pointwise at the test point. A sufficient condition is that the kernel satisfies $\|\phi(x) - \phi(y)\|_{\mathcal{H}} \le L_\phi \|x - y\|$ for some $L_\phi > 0$, in which case Assumption B.2 holds whenever

$$\frac{1}{T}\sum_{t=1}^{T} \|\hat{\varepsilon}_t - \varepsilon_t\|^2 \le (\delta_T / L_\phi)^2.$$

For the RBF kernel, $\|\phi(x) - \phi(y)\|_{\mathcal{H}}^2 = 2 - 2k(x,y) \le \|x - y\|^2/\ell^2$, so the condition holds with $L_\phi = 1/\ell$. The polynomial rate ρ at which the error vanishes depends on the estimation method and the problem structure. Faster estimators (larger ρ) yield tighter coverage guarantees.

Assumption B.3 (Boundedness)**.** *The kernel k is bounded: there exists $B > 0$ such that $k(x,x) \le B^2$ for all x in the domain.*

Remark. This is equivalent to $\|\phi(x)\|_{\mathcal{H}} \le B$ for all x. It ensures that all moments of $\phi(\varepsilon)$ exist ($\mathbb{E}[\|\phi(\varepsilon)\|^q] \le B^q$), which is required for the concentration inequalities used below.

Assumption B.4 (Trace-Class Covariance)**.** *The population covariance operator $\Sigma = \mathbb{E}[(\phi(\varepsilon) - \mu) \otimes (\phi(\varepsilon) - \mu)]$ is trace-class:* $\operatorname{tr}(\Sigma) < \infty$.

B.2 Coverage Guarantee and Convergence Rate (i.i.d.)

We prove the coverage guarantee through a sequence of lemmas, each bounding one source of error. These are combined in the proof of Theorem B.12.

B.2.1 Auxiliary Results

We collect standard results used throughout the proofs.

Lemma B.5 (Cauchy–Schwarz for Self-Adjoint Operators)**.** *For any self-adjoint operator A on $\mathcal{H}$ and any $x \in \mathcal{H}$,*

$$|\langle x, Ax\rangle_{\mathcal{H}}| \le \|x\|_{\mathcal{H}}^2 \, \|A\|_{\mathrm{op}}.$$

Lemma B.6 (Resolvent Identity)**.** *For invertible operators A and B on $\mathcal{H}$,*

$$A^{-1} - (A+B)^{-1} = A^{-1}B(A+B)^{-1}.$$

Proof. We have $A^{-1} = A^{-1}(A+B)(A+B)^{-1} = (I + A^{-1}B)(A+B)^{-1} = (A+B)^{-1} + A^{-1}B(A+B)^{-1}$. Rearranging gives the result. □

Lemma B.7 (McDiarmid's Inequality)**.** *Let $g : \mathcal{X}^m \to \mathbb{R}$ satisfy the bounded differences condition: for all $i \in \{1, \dots, m\}$, there exists $c_i < \infty$ such that*

$$\sup_{x_1,\dots,x_m,\tilde{x}_i} \big|g(x_1,\dots,x_m) - g(x_1,\dots,x_{i-1},\tilde{x}_i,x_{i+1},\dots,x_m)\big| \le c_i.$$

Then for independent random variables $X_1, \dots, X_m$ and any $\xi > 0$,

$$\mathbb{P}\big(g(X_1,\dots,X_m) - \mathbb{E}[g(X_1,\dots,X_m)] > \xi\big) \le \exp\!\left(-\frac{2\xi^2}{\sum_{i=1}^m c_i^2}\right).$$

B.2.2 Empirical CDF Concentration

Lemma B.8 (Convergence of Empirical CDF under i.i.d.)**.** *Under Assumption B.1, for any $T \ge 1$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{x\in\mathbb{R}} \big|\widetilde{F}_{T+1}(x) - F_e(x)\big| \le \sqrt{\frac{\log(2/\delta)}{2T}}. \tag{18}$$

Proof. Under Assumption B.1, the true scores $e_1, \dots, e_T$ are i.i.d. with CDF F_e. By the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, for any $\xi > 0$,

$$\mathbb{P}\Big(\sup_x \big|\widetilde{F}_{T+1}(x) - F_e(x)\big| > \xi\Big) \le 2\exp(-2T\xi^2).$$

Setting the right-hand side equal to δ and solving: $\delta = 2\exp(-2T\xi^2)$ gives $\xi = \sqrt{\log(2/\delta)/(2T)}$. □

B.2.3 Score Approximation Error

Lemma B.9 (Difference Between Estimated and True Nonconformity Scores)**.** *Under Assumptions B.1, B.2, B.3, and B.4, with probability at least $1-2\delta$,*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}|e_t - \hat{e}_t| \le \frac{4B\delta_T}{\lambda} + \frac{1}{\lambda}\big(4B\|\Delta_\mu\| + \|\Delta_\mu\|^2\big) \\ + \frac{4B^2}{\lambda^2}\Bigg[C_K\|\Sigma\|_{\mathrm{op}}\Bigg(\sqrt{\frac{\mathbf{r}(\Sigma)}{T}} \bigvee \frac{\mathbf{r}(\Sigma)}{T} \bigvee \sqrt{\frac{\log(1/\delta)}{T}} \bigvee \frac{\log(1/\delta)}{T}\Bigg) \\ + \left(\frac{2B}{\sqrt{T}} + B\sqrt{\frac{2\log(1/\delta)}{T}}\right)^2 + \frac{8B^2}{T-1} + \frac{T}{T-1}\big(8B\delta_T + 4\delta_T^2\big)\Bigg],\end{aligned} \tag{19}$$

where $\Delta_\mu = \mu - \hat{\mu}$ is the total mean estimation error, satisfying with probability at least $1-\delta$:

$$\|\Delta_\mu\| \le \delta_T + \frac{2B}{\sqrt{T}} + B\sqrt{\frac{2\log(1/\delta)}{T}} \tag{20}$$

and $C_K > 0$ is the constant from Koltchinskii et al. [18, Corollary 2].

Proof. We decompose the score difference via a telescoping argument. For each $t = 1, \ldots, T$, define the intermediate quantities:

$$I_1^{(t)} = \langle \hat{v}_t,\ \Sigma_\lambda^{-1} \hat{v}_t \rangle_{\mathcal{H}}$$
$$I_2^{(t)} = \langle \hat{w}_t,\ \Sigma_\lambda^{-1} \hat{w}_t \rangle_{\mathcal{H}}$$

where we use the notation:

$$\begin{aligned} v_t &= \phi(\varepsilon_t) - \mu && \text{(true features, population-centered)}, \\ \hat{v}_t &= \phi(\hat{\varepsilon}_t) - \mu && \text{(estimated features, population-centered)}, \\ \hat{w}_t &= \phi(\hat{\varepsilon}_t) - \hat{\mu} && \text{(estimated features, sample-centered)}. \end{aligned}$$

Note that $e_t = \langle v_t, \Sigma_\lambda^{-1} v_t \rangle$ and $\hat{e}_t = \langle \hat{w}_t, \widehat{\Sigma}_\lambda^{-1} \hat{w}_t \rangle$, and that $\hat{w}_t = \hat{v}_t + \Delta_\mu$ where $\Delta_\mu = \mu - \hat{\mu}$. By the triangle inequality:

$$|e_t - \hat{e}_t| \leq \underbrace{|e_t - I_1^{(t)}|}_{T_1^{(t)}\text{: residual error}} + \underbrace{|I_1^{(t)} - I_2^{(t)}|}_{T_2^{(t)}\text{: mean error}} + \underbrace{|I_2^{(t)} - \hat{e}_t|}_{T_3^{(t)}\text{: covariance error}}.$$

We bound each term separately.

Bounding $T_1^{(t)}$ (residual estimation error)

$$\begin{aligned} T_1^{(t)} &= \left| \langle v_t, \Sigma_\lambda^{-1} v_t \rangle - \langle \hat{v}_t, \Sigma_\lambda^{-1} \hat{v}_t \rangle \right| \\ &= \left| \langle v_t - \hat{v}_t,\ \Sigma_\lambda^{-1} v_t \rangle + \langle \hat{v}_t,\ \Sigma_\lambda^{-1} (v_t - \hat{v}_t) \rangle \right| \\ &= \left| \langle v_t - \hat{v}_t,\ \Sigma_\lambda^{-1} (v_t + \hat{v}_t) \rangle \right| \qquad \text{(self-adjointness of } \Sigma_\lambda^{-1}\text{)} \\ &\leq \|v_t - \hat{v}_t\| \cdot \|v_t + \hat{v}_t\| \cdot \|\Sigma_\lambda^{-1}\|_{\mathrm{op}} \qquad \text{(Lemma B.5)}. \end{aligned}$$

Since $\|\Sigma_\lambda^{-1}\|_{\mathrm{op}} \leq 1/\lambda$ and, by the triangle inequality and Assumption B.3,

$$\|v_t + \hat{v}_t\| = \|\phi(\varepsilon_t) + \phi(\hat{\varepsilon}_t) - 2\mu\| \leq \|\phi(\varepsilon_t)\| + \|\phi(\hat{\varepsilon}_t)\| + 2\|\mu\| \leq 4B,$$

we obtain

$$T_1^{(t)} \leq \frac{4B}{\lambda} \|\phi(\varepsilon_t) - \phi(\hat{\varepsilon}_t)\|.$$

Summing over t and applying the Cauchy–Schwarz inequality:

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} T_1^{(t)} &\leq \frac{4B}{\lambda} \cdot \frac{1}{T} \sum_{t=1}^{T} \|\phi(\varepsilon_t) - \phi(\hat{\varepsilon}_t)\| \\ &\leq \frac{4B}{\lambda} \sqrt{\frac{1}{T} \sum_{t=1}^{T} \|\phi(\varepsilon_t) - \phi(\hat{\varepsilon}_t)\|^2} \\ &\leq \frac{4B\delta_T}{\lambda}, \end{aligned}$$

where the last step uses Assumption B.2.

Bounding $T_2^{(t)}$ (mean estimation error). Since $\hat{w}_t = \hat{v}_t + \Delta_\mu$:

$$\begin{aligned} T_2^{(t)} &= \left| \langle \hat{v}_t, \Sigma_\lambda^{-1} \hat{v}_t \rangle - \langle \hat{v}_t + \Delta_\mu, \Sigma_\lambda^{-1} (\hat{v}_t + \Delta_\mu) \rangle \right| \\ &= \left| \langle \hat{v}_t, \Sigma_\lambda^{-1} \hat{v}_t \rangle - \langle \hat{v}_t, \Sigma_\lambda^{-1} \hat{v}_t \rangle - 2\langle \hat{v}_t, \Sigma_\lambda^{-1} \Delta_\mu \rangle - \langle \Delta_\mu, \Sigma_\lambda^{-1} \Delta_\mu \rangle \right| \\ &= \left| 2\langle \hat{v}_t, \Sigma_\lambda^{-1} \Delta_\mu \rangle + \langle \Delta_\mu, \Sigma_\lambda^{-1} \Delta_\mu \rangle \right| \\ &\leq 2 \left| \langle \hat{v}_t, \Sigma_\lambda^{-1} \Delta_\mu \rangle \right| + \left| \langle \Delta_\mu, \Sigma_\lambda^{-1} \Delta_\mu \rangle \right|. \end{aligned}$$

By Cauchy–Schwarz and $\|\hat{v}_t\| = \|\phi(\hat{\varepsilon}_t) - \mu\| \le 2B$:

$$2\big|\langle \hat{v}_t, \Sigma_\lambda^{-1}\Delta_\mu\rangle\big| \le 2\|\hat{v}_t\| \cdot \|\Sigma_\lambda^{-1}\|_{\mathrm{op}} \cdot \|\Delta_\mu\| \le \frac{4B}{\lambda}\|\Delta_\mu\|,$$
$$\big|\langle \Delta_\mu, \Sigma_\lambda^{-1}\Delta_\mu\rangle\big| \le \|\Delta_\mu\|^2 \cdot \|\Sigma_\lambda^{-1}\|_{\mathrm{op}} \le \frac{1}{\lambda}\|\Delta_\mu\|^2$$

Since this bound is independent of t:

$$\frac{1}{T}\sum_{t=1}^{T} T_2^{(t)} \le \frac{1}{\lambda}\big(4B\|\Delta_\mu\| + \|\Delta_\mu\|^2\big).$$

We now bound $\|\Delta_\mu\| = \|\mu - \hat{\mu}\|$. Introducing the sample mean of true features $\tilde{\mu} = (1/T)\sum_{t=1}^{T}\phi(\varepsilon_t)$:

$$\|\Delta_\mu\| = \|\mu - \hat{\mu}\| \le \|\mu - \tilde{\mu}\| + \|\tilde{\mu} - \hat{\mu}\|.$$

The second term is bounded by Assumption B.2:

$$\|\tilde{\mu} - \hat{\mu}\| = \Big\|\frac{1}{T}\sum_{t=1}^{T}\big(\phi(\varepsilon_t) - \phi(\hat{\varepsilon}_t)\big)\Big\|$$
$$\le \sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\phi(\varepsilon_t) - \phi(\hat{\varepsilon}_t)\|^2} \le \delta_T.$$

For $\|\mu - \tilde{\mu}\|$, we apply McDiarmid's inequality (Lemma B.7). Define

$$g(\varepsilon_1, \ldots, \varepsilon_T) = \|\tilde{\mu} - \mu\| = \Big\|\frac{1}{T}\sum_{t=1}^{T}\phi(\varepsilon_t) - \mu\Big\|.$$

Replacing any single ε_i changes g by at most

$$|g(\ldots, \varepsilon_i, \ldots) - g(\ldots, \tilde{\varepsilon}_i, \ldots)| \le \frac{1}{T}\|\phi(\varepsilon_i) - \phi(\tilde{\varepsilon}_i)\| \le \frac{2B}{T}$$

so the bounded differences condition holds with $c_i = 2B/T$ for all i. By Lemma B.7:

$$\mathbb{P}\big(\|\tilde{\mu} - \mu\| - \mathbb{E}\|\tilde{\mu} - \mu\| > \xi\big) \le \exp\!\left(-\frac{2\xi^2}{T \cdot (2B/T)^2}\right) = \exp\!\left(-\frac{T\xi^2}{2B^2}\right).$$

We bound the expectation using the i.i.d. assumption:

$$\mathbb{E}\|\tilde{\mu} - \mu\|^2 = \frac{1}{T^2}\sum_{t_1,t_2}\mathbb{E}\big[\langle \phi(\varepsilon_{t_1}) - \mu, \phi(\varepsilon_{t_2}) - \mu\rangle\big]$$
$$= \frac{1}{T}\,\mathbb{E}\|\phi(\varepsilon) - \mu\|^2 \qquad \text{(cross terms vanish by independence)}$$
$$\le \frac{4B^2}{T}$$

where the last step uses $\|\phi(\varepsilon) - \mu\|^2 \le 2(\|\phi(\varepsilon)\|^2 + \|\mu\|^2) \le 4B^2$. By Jensen's inequality:

$$\mathbb{E}\|\tilde{\mu} - \mu\| \le \sqrt{\mathbb{E}\|\tilde{\mu} - \mu\|^2} \le \frac{2B}{\sqrt{T}}.$$

Setting $\delta = \exp(-T\xi^2/(2B^2))$ and solving for ξ: with probability at least $1 - \delta$,

$$\|\tilde{\mu} - \mu\| \le \frac{2B}{\sqrt{T}} + B\sqrt{\frac{2\log(1/\delta)}{T}}.$$

Combining with $\|\tilde{\mu} - \hat{\mu}\| \le \delta_T$: with probability at least $1 - \delta$,

$$\|\Delta_\mu\| \le \delta_T + \frac{2B}{\sqrt{T}} + B\sqrt{\frac{2\log(1/\delta)}{T}}. \tag{21}$$

Bounding $T_3^{(t)}$ (covariance estimation error). Recall $\hat{w}_t = \phi(\hat{\varepsilon}_t) - \hat{\mu}$, so $\|\hat{w}_t\| \le 2B$. By Lemma B.5:

$$\begin{aligned} T_3^{(t)} &= \left|\langle \hat{w}_t, \Sigma_\lambda^{-1}\hat{w}_t\rangle - \langle \hat{w}_t, \widehat{\Sigma}_\lambda^{-1}\hat{w}_t\rangle\right| \\ &= \left|\langle \hat{w}_t, (\Sigma_\lambda^{-1} - \widehat{\Sigma}_\lambda^{-1})\hat{w}_t\rangle\right| \\ &\le \|\hat{w}_t\|^2 \cdot \|\Sigma_\lambda^{-1} - \widehat{\Sigma}_\lambda^{-1}\|_{\mathrm{op}} \\ &\le 4B^2 \cdot \|\Sigma_\lambda^{-1} - \widehat{\Sigma}_\lambda^{-1}\|_{\mathrm{op}}. \end{aligned}$$

By the resolvent identity (Lemma B.6) with $A = \Sigma + \lambda I$ and $B = \widehat{\Sigma} - \Sigma$:

$$\begin{aligned} \|\Sigma_\lambda^{-1} - \widehat{\Sigma}_\lambda^{-1}\|_{\mathrm{op}} &= \|(\Sigma + \lambda I)^{-1}(\widehat{\Sigma} - \Sigma)(\widehat{\Sigma} + \lambda I)^{-1}\|_{\mathrm{op}} \\ &\le \|(\Sigma + \lambda I)^{-1}\|_{\mathrm{op}} \cdot \|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \cdot \|(\widehat{\Sigma} + \lambda I)^{-1}\|_{\mathrm{op}} \\ &\le \frac{1}{\lambda^2}\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}. \end{aligned}$$

We decompose $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \le \|\widetilde{\Sigma} - \Sigma\|_{\mathrm{op}} + \|\widehat{\Sigma} - \widetilde{\Sigma}\|_{\mathrm{op}}$ and bound each part.

Statistical error $\|\widetilde{\Sigma} - \Sigma\|_{\mathrm{op}}$. Define $v_t = \phi(\varepsilon_t) - \mu$, $\Delta = \tilde{\mu} - \mu = (1/T)\sum_{t=1}^T v_t$, and $C_T = (1/T)\sum_{t=1}^T v_t \otimes v_t$.

Since $\widetilde{\Sigma} = \frac{1}{T-1}\sum_{t=1}^T(\phi(\varepsilon_t) - \tilde{\mu}) \otimes (\phi(\varepsilon_t) - \tilde{\mu})$ and $\phi(\varepsilon_t) - \tilde{\mu} = v_t - \Delta$, a direct computation gives:

$$\begin{aligned} \widetilde{\Sigma} &= \frac{T}{T-1} \cdot \frac{1}{T}\sum_{t=1}^T (v_t - \Delta) \otimes (v_t - \Delta) \\ &= \frac{T}{T-1}\left[\frac{1}{T}\sum_{t=1}^T v_t \otimes v_t - \Delta \otimes \Delta\right] = \frac{T}{T-1}\big[C_T - \Delta \otimes \Delta\big]. \end{aligned}$$

Therefore:

$$\begin{aligned} \widetilde{\Sigma} - \Sigma &= \frac{T}{T-1}(C_T - \Delta \otimes \Delta) - \Sigma \\ &= (C_T - \Sigma) + \frac{1}{T-1}C_T - \frac{T}{T-1}(\Delta \otimes \Delta). \end{aligned}$$

Taking operator norms and using $\frac{T}{T-1}\|\Delta\|^2 = \|\Delta\|^2 + \frac{1}{T-1}\|\Delta\|^2$:

$$\|\widetilde{\Sigma} - \Sigma\|_{\mathrm{op}} \le \|C_T - \Sigma\|_{\mathrm{op}} + \|\Delta\|^2 + \frac{1}{T-1}\|C_T\|_{\mathrm{op}} + \frac{1}{T-1}\|\Delta\|^2.$$

Since $\|v_t\| \le 2B$ (Assumption B.3):

$$\|C_T\|_{\mathrm{op}} \le \frac{1}{T}\sum_{t=1}^T \|v_t\|^2 \le 4B^2 \qquad \|\Delta\| \le \frac{1}{T}\sum_{t=1}^T \|v_t\| \le 2B$$

giving $\frac{1}{T-1}\|C_T\|_{\mathrm{op}} \le \frac{4B^2}{T-1}$ and $\frac{1}{T-1}\|\Delta\|^2 \le \frac{4B^2}{T-1}$.

For $\|C_T - \Sigma\|_{\mathrm{op}}$: since $\|v_t\| \le 2B$, the random vectors v_t are bounded, hence subgaussian. Combined with $\mathrm{tr}(\Sigma) < \infty$ (Assumption B.4), the v_t are pregaussian, satisfying the conditions of [18, Corollary 2]. For any $\delta \in (0,1)$, with probability at least $1 - \delta$:

$$\|C_T - \Sigma\|_{\mathrm{op}} \le C_K\|\Sigma\|_{\mathrm{op}}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{T}} \bigvee \frac{\mathbf{r}(\Sigma)}{T} \bigvee \sqrt{\frac{\log(1/\delta)}{T}} \bigvee \frac{\log(1/\delta)}{T}\right) \tag{22}$$

where $\mathbf{r}(\Sigma) = \mathbb{E}\|v_t\|^2/\|\Sigma\|_{\mathrm{op}} \le 4B^2/\|\Sigma\|_{\mathrm{op}}$.

Using the bound on $\|\Delta\|^2$ from the mean estimation analysis (which holds with probability at least $1-\delta$, a union bound over the covariance and mean events gives: with probability at least $1-2\delta$,

$$\begin{aligned}\|\widetilde{\Sigma}-\Sigma\|_{\mathrm{op}} \le C_K\|\Sigma\|_{\mathrm{op}}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{T}}\bigvee\frac{\mathbf{r}(\Sigma)}{T}\bigvee\sqrt{\frac{\log(1/\delta)}{T}}\bigvee\frac{\log(1/\delta)}{T}\Big)\\ +\Big(\frac{2B}{\sqrt{T}}+B\sqrt{\frac{2\log(1/\delta)}{T}}\Big)^2+\frac{8B^2}{T-1}.\end{aligned} \tag{23}$$

Prediction error $\|\widehat{\Sigma}-\widetilde{\Sigma}\|_{\mathrm{op}}$. Define $\tilde{w}_t = \phi(\varepsilon_t)-\tilde{\mu}$ (true features, sample-centered), $\Delta_t = \phi(\hat{\varepsilon}_t)-\phi(\varepsilon_t)$ (a pointwise prediction error), and $\bar{\Delta}=(1/T)\sum_{t=1}^T\Delta_t=\hat{\mu}-\tilde{\mu}$ (the mean prediction error).

Then $\hat{w}_t=\phi(\hat{\varepsilon}_t)-\hat{\mu}=(\phi(\varepsilon_t)+\Delta_t)-(\tilde{\mu}+\bar{\Delta})=\tilde{w}_t+\theta_t$, where $\theta_t=\Delta_t-\bar{\Delta}$ is the demeaned prediction error. Expanding:

$$\begin{aligned}\|\widehat{\Sigma}-\widetilde{\Sigma}\|_{\mathrm{op}} &= \Big\|\frac{1}{T-1}\sum_{t=1}^T\big(\hat{w}_t\otimes\hat{w}_t-\tilde{w}_t\otimes\tilde{w}_t\big)\Big\|_{\mathrm{op}}\\ &= \Big\|\frac{1}{T-1}\sum_{t=1}^T\big(\tilde{w}_t\otimes\theta_t+\theta_t\otimes\tilde{w}_t+\theta_t\otimes\theta_t\big)\Big\|_{\mathrm{op}}\\ &\le \frac{1}{T-1}\sum_{t=1}^T\big(2\|\tilde{w}_t\|\,\|\theta_t\|+\|\theta_t\|^2\big).\end{aligned}$$

We have $\|\tilde{w}_t\|=\|\phi(\varepsilon_t)-\tilde{\mu}\|\le 2B$ and $\|\theta_t\|=\|\Delta_t-\bar{\Delta}\|\le\|\Delta_t\|+\|\bar{\Delta}\|$. By Assumption B.2:

$$\|\bar{\Delta}\|=\Big\|\frac{1}{T}\sum_{t=1}^T\Delta_t\Big\|\le\sqrt{\frac{1}{T}\sum_{t=1}^T\|\Delta_t\|^2}\le\delta_T.$$

Using $(1/T)\sum_t\|\Delta_t\|\le\delta_T$ and $(1/T)\sum_t\|\Delta_t\|^2\le\delta_T^2$:

$$\begin{aligned}\|\widehat{\Sigma}-\widetilde{\Sigma}\|_{\mathrm{op}} &\le \frac{T}{T-1}\cdot\frac{1}{T}\sum_{t=1}^T\big(4B(\|\Delta_t\|+\delta_T)+(\|\Delta_t\|+\delta_T)^2\big)\\ &\le \frac{T}{T-1}\big(8B\delta_T+4\delta_T^2\big).\end{aligned}$$

Combining. Summing and dividing by T:

$$\frac{1}{T}\sum_{t=1}^T T_3^{(t)}\le\frac{4B^2}{\lambda^2}\Big[\|\widetilde{\Sigma}-\Sigma\|_{\mathrm{op}}+\|\widehat{\Sigma}-\widetilde{\Sigma}\|_{\mathrm{op}}\Big].$$

The bounds on $\|\widetilde{\Sigma}-\Sigma\|_{\mathrm{op}}$ (Equation (23)) and $\|\widehat{\Sigma}-\widetilde{\Sigma}\|_{\mathrm{op}}$ hold simultaneously with probability at least $1-2\delta$. Adding $T_1+T_2+T_3$ gives (19). □

B.2.4 ECDF Coupling

Lemma B.10 (Distance Between Empirical CDFs of True and Estimated Scores)**.** *Under Assumptions B.1, B.2, B.3, and B.4, with probability at least* $1-2\delta$*,*

$$\sup_{x\in\mathbb{R}}\big|\widehat{F}_{T+1}(x)-\widetilde{F}_{T+1}(x)\big|\le(L+1)\,C_S+2\sup_{x\in\mathbb{R}}\big|\widetilde{F}_{T+1}(x)-F_e(x)\big| \tag{24}$$

where $C_S=\big[(1/T)\sum_{t=1}^T|e_t-\hat{e}_t|\big]^{1/2}$ *and* L *is the Lipschitz constant of* F_e*.*

Proof. The argument follows [40, Lemma 4.11]. For any threshold x:

$$\big|\widehat{F}_{T+1}(x) - \widetilde{F}_{T+1}(x)\big| = \frac{1}{T}\Big|\sum_{t=1}^{T}\big(\mathbb{1}_{\{\hat{e}_t \le x\}} - \mathbb{1}_{\{e_t \le x\}}\big)\Big| \le \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\{|e_t - x| \le |e_t - \hat{e}_t|\}}.$$

For any $\eta > 0$, we split each indicator:

$$\mathbb{1}_{\{|e_t - x| \le |e_t - \hat{e}_t|\}} \le \mathbb{1}_{\{|e_t - x| \le \eta\}} + \mathbb{1}_{\{|e_t - \hat{e}_t| > \eta\}}.$$

For the second term, by Markov's inequality:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\{|e_t - \hat{e}_t| > \eta\}} \le \frac{1}{\eta}\cdot\frac{1}{T}\sum_{t=1}^{T}|e_t - \hat{e}_t| = \frac{C_S^2}{\eta}.$$

For the first term, using the Lipschitz property of F_e:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\{|e_t - x| \le \eta\}} &\le \widetilde{F}_{T+1}(x+\eta) - \widetilde{F}_{T+1}(x-\eta) \\ &\le F_e(x+\eta) - F_e(x-\eta) + 2\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \\ &\le 2L\eta + 2\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)|.\end{aligned}$$

Combining, $|\widehat{F}_{T+1}(x) - \widetilde{F}_{T+1}(x)| \le 2L\eta + C_S^2/\eta + 2\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)|$. Treating the two signs of $\widehat{F}_{T+1}(x) - \widetilde{F}_{T+1}(x)$ separately replaces $2L\eta$ by $L\eta$, and $\eta = C_S$ gives the stated bound [40, Lemma 4.11]. □

B.2.5 Test Point Score Error

Lemma B.11 (Test Point Nonconformity Score Error)**.** *Under Assumptions B.1, B.2, B.3, and B.4, with probability at least* $1 - 2\delta$*,*

$$|e_{T+1} - \hat{e}_{T+1}| \le \zeta_T \tag{25}$$

where ζ_T *has the same form as the right-hand side of* (19)*.*

Proof. By the same telescoping decomposition:

$$|e_{T+1} - \hat{e}_{T+1}| \le T_1^{(T+1)} + T_2^{(T+1)} + T_3^{(T+1)}.$$

The bounds on $T_2^{(T+1)}$ and $T_3^{(T+1)}$ are identical to those in Lemma B.9, since they depend only on $\hat{\mu}$, $\widehat{\Sigma}$, and Σ (shared between calibration and test point). For $T_1^{(T+1)}$, the same derivation gives:

$$T_1^{(T+1)} \le \frac{4B}{\lambda}\|\phi(\varepsilon_{T+1}) - \phi(\hat{\varepsilon}_{T+1})\| \le \frac{4B\delta_T}{\lambda}$$

using the test-point bound in Assumption B.2. Summing all three terms gives $|e_{T+1} - \hat{e}_{T+1}| \le \zeta_T$, where ζ_T has the same expression as $(1/T)\sum_t |e_t - \hat{e}_t| = C_S^2$ from Lemma B.9. □

B.2.6 Coverage Guarantee

Theorem B.12 (Asymptotic Conditional Coverage Guarantee under i.i.d.)**.** *Under Assumptions B.1, B.2, B.3, and B.4, for any* $T \ge 1$*,* $\alpha \in (0,1)$*, and* $\delta \in (0,1)$*, with probability at least* $1 - 4\delta$*:*

$$\big|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} \mid X_{T+1}) - (1-\alpha)\big| \le 12\sqrt{\frac{\log(2/\delta)}{2T}} + 4(L+1)(C_S + \zeta_T) + 2\delta.$$

Proof. Under Assumption B.1, $\varepsilon_{T+1} \perp X_{T+1}$, so e_{T+1} is a function of ε_{T+1} alone and satisfies $F_e(e_{T+1}) \mid X_{T+1} \sim \mathrm{Uniform}(0,1)$. Furthermore, conditioned on $X_{T+1} = x$, the estimated score $\hat{e}_{T+1} = \hat{e}_k(f(x) - \hat{f}(x) + \varepsilon_{T+1})$ depends on x only through the deterministic shift $f(x) - \hat{f}(x)$, with all remaining randomness from $\varepsilon_{T+1} \perp x$. All subsequent bounds therefore hold conditionally on any fixed $X_{T+1} = x$, and the $\mid X_{T+1}$ notation is justified throughout this proof.

For any $\beta \in [0, \alpha]$, by construction of $\widehat{C}^{\alpha}_{T+1}$:

$$Y_{T+1} \in \widehat{C}^{\alpha}_{T+1} \iff \beta \le \widehat{F}_{T+1}(\hat{e}_{T+1}) \le 1 - \alpha + \beta$$

Since $F_e(e_{T+1}) \sim \mathrm{Uniform}(0,1)$ under Assumption B.1, by the elementary bound $|\mathbb{1}\{a \le x \le b\} - \mathbb{1}\{a \le y \le b\}| \le |\mathbb{1}\{a \le x\} - \mathbb{1}\{a \le y\}| + |\mathbb{1}\{x \le b\} - \mathbb{1}\{y \le b\}|$:

$$\begin{aligned}
&\big|\mathbb{P}(Y_{T+1} \in \widehat{C}^{\alpha}_{T+1} \mid X_{T+1}) - (1-\alpha)\big| \\
&\le \mathbb{P}\big(|F_e(e_{T+1}) - \beta| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})|\big) \\
&\quad + \mathbb{P}\big(|F_e(e_{T+1}) - (1-\alpha+\beta)| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})|\big)
\end{aligned}$$

Let A_T denote the event in Lemma B.8: $\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \le \sqrt{\log(2/\delta)/(2T)}$, with $\mathbb{P}(A_T) \ge 1 - \delta$. For any $\gamma \in [0,1]$:

$$\begin{aligned}
&\mathbb{P}\big(|F_e(e_{T+1}) - \gamma| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})|\big) \\
&\le \mathbb{P}\big(|F_e(e_{T+1}) - \gamma| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(\hat{e}_{T+1})| + |F_e(\hat{e}_{T+1}) - F_e(e_{T+1})| \;\big|\; A_T\big) + \delta
\end{aligned}$$

Conditioning on A_T, we bound $|\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(\hat{e}_{T+1})|$ by the triangle inequality:

$$|\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(\hat{e}_{T+1})| \le \sup_x |\widehat{F}_{T+1}(x) - \widetilde{F}_{T+1}(x)| + \sup_x |\widetilde{F}_{T+1}(x) - F_e(x)|$$

By Lemma B.10 and the definition of A_T:

$$\sup_x |\widehat{F}_{T+1}(x) - \widetilde{F}_{T+1}(x)| \le (L+1)C_S + 2\sqrt{\frac{\log(2/\delta)}{2T}}$$

Adding $\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \le \sqrt{\log(2/\delta)/(2T)}$ on A_T:

$$|\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(\hat{e}_{T+1})| \le (L+1)C_S + 3\sqrt{\frac{\log(2/\delta)}{2T}}$$

By the Lipschitz continuity of F_e and Lemma B.11:

$$|F_e(\hat{e}_{T+1}) - F_e(e_{T+1})| \le L|\hat{e}_{T+1} - e_{T+1}| \le L\zeta_T$$

Combining:

$$\begin{aligned}
|\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})| &\le (L+1)C_S + 3\sqrt{\frac{\log(2/\delta)}{2T}} + L\zeta_T \\
&\le 3\sqrt{\frac{\log(2/\delta)}{2T}} + (L+1)(C_S + \zeta_T)
\end{aligned}$$

Since $F_e(e_{T+1}) \sim \mathrm{Uniform}(0,1)$, $\mathbb{P}(|F_e(e_{T+1}) - \gamma| \le \eta \mid A_T) \le 2\eta$. Therefore:

$$\begin{aligned}
&\mathbb{P}\big(|F_e(e_{T+1}) - \gamma| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})| \;\big|\; A_T\big) \\
&\le 6\sqrt{\frac{\log(2/\delta)}{2T}} + 2(L+1)(C_S + \zeta_T)
\end{aligned}$$

Adding $\mathbb{P}(A_T^c) \le \delta$ and applying with $\gamma = \beta$ and $\gamma = 1 - \alpha + \beta$:

$$\begin{aligned}
&\left|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} \mid X_{T+1}) - (1-\alpha)\right| \\
&\le 2\Big[6\sqrt{\frac{\log(2/\delta)}{2T}} + 2(L+1)(C_S + \zeta_T) + \delta\Big] \\
&= 12\sqrt{\frac{\log(2/\delta)}{2T}} + 4(L+1)(C_S + \zeta_T) + 2\delta
\end{aligned}$$

The probability is at least $1 - 4\delta$ by a union bound over: DKW $(1-\delta)$, mean concentration $(1-\delta)$, covariance concentration $(1-\delta)$, and test-point estimation $(1-\delta)$. □

B.2.7 Convergence Rate

Theorem B.13 (Conditional Coverage Convergence Rate under i.i.d.)**.** *Under Assumptions B.1–B.4, there exists a constant $C > 0$ depending only on B, C_K, $\|\Sigma\|_{\mathrm{op}}$, L such that with probability at least $1 - 4T^{-1}$:*

$$\left|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} \mid X_{T+1}) - (1-\alpha)\right| \le \begin{cases} C \cdot T^{\beta - 1/4}(\log T)^{1/4} & \text{if } \rho \ge 1/2,\ \lambda = T^{-\beta},\ \beta \in (0, 1/4) \\ C \cdot T^{\beta - \rho/2} & \text{if } \rho < 1/2,\ \lambda = T^{-\beta},\ \beta \in (0, \rho/2) \end{cases}$$

Proof. Set $\delta = T^{-1}$ (so $\log(1/\delta) = \log T$) and $\lambda = T^{-\beta}$. We determine the asymptotic rate of each term in Lemma B.9.

Residual error term.

$$\frac{4B\delta_T}{\lambda} = O(T^{\beta - \rho}) \tag{26}$$

Mean error term. From (21):

$$\|\Delta_\mu\| \le \delta_T + \frac{2B}{\sqrt{T}} + B\sqrt{\frac{2\log T}{T}} = O\big(\delta_T + T^{-1/2}\sqrt{\log T}\big)$$

If $\rho \ge 1/2$, the $T^{-1/2}\sqrt{\log T}$ term dominates $\delta_T = O(T^{-\rho})$, so $\|\Delta_\mu\| = O(T^{-1/2}\sqrt{\log T})$. Since $\|\Delta_\mu\|^2 = O(T^{-1}\log T)$ is dominated by $\|\Delta_\mu\|$ for large T:

$$\frac{1}{\lambda}\big(4B\|\Delta_\mu\| + \|\Delta_\mu\|^2\big) = O\big(T^{\beta - 1/2}\sqrt{\log T}\big)$$

If $\rho < 1/2$, then δ_T dominates and:

$$\frac{1}{\lambda}\big(4B\|\Delta_\mu\| + \|\Delta_\mu\|^2\big) = O(T^{\beta - \rho})$$

In both cases, this term vanishes for $\beta < \min(\rho, 1/2)$. We write:

$$\frac{1}{\lambda}\big(4B\|\Delta_\mu\| + \|\Delta_\mu\|^2\big) = O\big(T^{\beta - \min(\rho, 1/2)}\sqrt{\log T}\big) \tag{27}$$

Covariance error term. Define the effective rank bound $r_0 = 4B^2/\|\Sigma\|_{\mathrm{op}} \ge \mathbf{r}(\Sigma)$. For $T \ge e^{r_0}$, we have $\log T \ge r_0 \ge \mathbf{r}(\Sigma)$, so:

$$\sqrt{\frac{\mathbf{r}(\Sigma)}{T}} \le \sqrt{\frac{\log T}{T}}, \qquad \frac{\mathbf{r}(\Sigma)}{T} \le \frac{\log T}{T}$$

Since $\sqrt{\log T/T}$ dominates $\log T/T$ for all T, the Koltchinskii bound (22) simplifies to:

$$\|C_T - \Sigma\|_{\mathrm{op}} = O\left(\sqrt{\frac{\log T}{T}}\right) \tag{28}$$

The remaining terms in (23):

$$\left(\frac{2B}{\sqrt{T}} + B\sqrt{\frac{2\log T}{T}}\right)^2 = O\left(\frac{\log T}{T}\right) \tag{29}$$

$$\frac{8B^2}{T-1} = O(T^{-1}), \qquad \frac{T}{T-1}\left(8B\delta_T + 4\delta_T^2\right) = O(\delta_T) = O(T^{-\rho}). \tag{30}$$

Since $O(\sqrt{\log T/T})$ dominates both $O(\log T/T)$ and $O(T^{-1})$, and $O(T^{-\rho})$ dominates $O(T^{-1})$ for $\rho < 1$, we denote the full covariance error bracket by A:

$$A = \|\widetilde{\Sigma} - \Sigma\|_{\mathrm{op}} + \|\widehat{\Sigma} - \widetilde{\Sigma}\|_{\mathrm{op}} = O\left(\sqrt{\frac{\log T}{T}}\right) + O(T^{-\rho}).$$

Multiplying by $4B^2/\lambda^2 = O(T^{2\beta})$:

$$\frac{4B^2}{\lambda^2} \cdot A = O(T^{2\beta}) \cdot \left[O\left(T^{-1/2}\sqrt{\log T}\right) + O(T^{-\rho})\right]. \tag{31}$$

For $\rho \geq 1/2$, the first term dominates:

$$\frac{4B^2}{\lambda^2} \cdot A = O(T^{2\beta - 1/2}\sqrt{\log T})$$

which vanishes for $\beta < 1/4$. For $\rho < 1/2$, the second term dominates:

$$\frac{4B^2}{\lambda^2} \cdot A = O(T^{2\beta - \rho})$$

which vanishes for $\beta < \rho/2$.

Dominant term. Comparing the exponents from (26), (27), and (31):

$$\beta - \rho \;\leq\; \beta - \min(\rho, 1/2) \;<\; 2\beta - \min(\rho, 1/2)$$

so the covariance term (31) dominates. Thus:

$$C_S^2 = \zeta_T = \begin{cases} O(T^{2\beta - 1/2}\sqrt{\log T}) & \text{if } \rho \geq 1/2, \\ O(T^{2\beta - \rho}) & \text{if } \rho < 1/2 \end{cases} \tag{32}$$

where we use that C_S^2 is the calibration average bound from Lemma B.9 and ζ_T is the test-point bound from Lemma B.11.

Taking square roots: $C_S = O(T^{\beta - 1/4}(\log T)^{1/4})$ for $\rho \geq 1/2$, and $C_S = O(T^{\beta - \rho/2})$ for $\rho < 1/2$.

Final bound. From Theorem B.12 with $\delta = T^{-1}$, with probability at least $1 - 4T^{-1}$:

$$\left|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} \mid X_{T+1}) - (1-\alpha)\right| \leq \underbrace{12\sqrt{\frac{\log(2T)}{2T}}}_{O(T^{-1/2}\sqrt{\log T})} + \underbrace{4(L+1)(C_S + \zeta_T)}_{\text{score error}} + \underbrace{2T^{-1}}_{O(T^{-1})}.$$

Since $\zeta_T = C_S^2$ and $C_S \to 0$ (for $\beta < \min(\rho/2, 1/4)$), we have $C_S + \zeta_T \leq 2C_S$ for large T. The C_S term dominates the DKW and T^{-1} terms because $\beta - \min(\rho/2, 1/4) > -1/2 > -1$ for $\beta, \rho > 0$, giving:

$$\left|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} \mid X_{T+1}) - (1-\alpha)\right| \leq \begin{cases} C \cdot T^{\beta - 1/4}(\log T)^{1/4} & \text{if } \rho \geq 1/2 \\ C \cdot T^{\beta - \rho/2} & \text{if } \rho < 1/2 \end{cases}$$

where C depends on B, C_K, $\|\Sigma\|_{\mathrm{op}}$, L. □

Remark. The constant C depends on B, C_K, $\|\Sigma\|_{\mathrm{op}}$ and L, and its dependence on the kernel can be characterised for common choices. Any stationary kernel with $k(0) = 1$, such as the RBF, Matérn and Laplace kernels, has $B = 1$. Since Σ is built from centered features, $\mathrm{tr}(\Sigma) = 1 - \|\mu\|_{\mathcal{H}}^2 \leq 1$ in every dimension. The spectral factor of the covariance concentration bound then satisfies $\|\Sigma\|_{\mathrm{op}}\sqrt{\mathbf{r}(\Sigma)} = \sqrt{\mathrm{tr}(\Sigma)\,\|\Sigma\|_{\mathrm{op}}} \leq 1$. For these kernels, the part of the constant that depends on the kernel is therefore bounded by universal numbers that do not depend on d. The only remaining quantity is the Lipschitz constant L, which is a property of the residual distribution. For comparison, the linear kernel on standardized and bounded residuals gives $\mathrm{tr}(\Sigma) = d$, which recovers the dimension dependence of the Mahalanobis score. The kernel also enters the exponent through ρ in Assumption B.2, since δ_T measures the error of the base model in the RKHS norm of the kernel.

B.3 Coverage under α-Mixing Dependence

We extend the coverage analysis to sequences with temporal dependence. We assume that the mean and covariance operator are known (Assumption B.15), eliminating the mean and covariance estimation error and isolating the effect of mixing on conformal calibration.

B.3.1 Additional Assumptions

Assumption B.14 (Stationary Strong Mixing)**.** *The process $\{\varepsilon_t\}_{t\geq 1}$ is stationary and α-mixing with mixing coefficients $\{\alpha_k\}_{k\geq 0}$, satisfying $\sum_{k=0}^{\infty}\alpha_k \leq M < \infty$. Furthermore, the CDF F_e of the true nonconformity score is L-Lipschitz.*

Assumption B.15 (Known Mean and Covariance)**.** *The population mean $\mu = \mathbb{E}[\phi(\varepsilon)]$ and covariance operator $\Sigma = \mathbb{E}[(\phi(\varepsilon) - \mu) \otimes (\phi(\varepsilon) - \mu)]$ are known.*

Remark. Under Assumption B.15, we set $\hat{\mu} = \mu$ and $\widehat{\Sigma} = \Sigma$ in the construction of the nonconformity scores for analysis. This eliminates the mean estimation error (term T_2) and the covariance estimation error (term T_3) from the decomposition in Lemma B.9, leaving only the residual estimation error (term T_1).

This assumption isolates the effect of temporal dependence on conformal calibration. Analyzing covariance operator estimation under α-mixing in infinite-dimensional RKHSs would require concentration inequalities considerably more involved than their i.i.d. counterparts and is beyond the scope of this work. In practice, Σ can be estimated from an independent sample or from a sufficiently long stationary pre-training period.

B.3.2 Inheritance of Mixing

Lemma B.16 (Inheritance of Mixing by Nonconformity Scores)**.** *Let $\{\varepsilon_t\}_{t\geq 1}$ be a stationary α-mixing sequence with mixing coefficients $\{\alpha_k\}_{k\geq 0}$. Define*

$$e_t = \langle \phi(\varepsilon_t) - \mu,\ \Sigma_\lambda^{-1}(\phi(\varepsilon_t) - \mu)\rangle_{\mathcal{H}}.$$

Then $\{e_t\}_{t\geq 1}$ is stationary and α-mixing with $\alpha_k(\{e_t\}) \leq \alpha_k(\{\varepsilon_t\})$ for all $k \geq 0$.

Proof. The map $\varepsilon \mapsto e(\varepsilon) = \langle \phi(\varepsilon) - \mu, \Sigma_\lambda^{-1}(\phi(\varepsilon) - \mu)\rangle_{\mathcal{H}}$ is measurable as a composition of continuous functions: ϕ is continuous (for RBF, Matérn, polynomial kernels) and $v \mapsto \langle v, \Sigma_\lambda^{-1} v\rangle$ is continuous on $\mathcal{H}$ since $\|\Sigma_\lambda^{-1}\|_{\mathrm{op}} \leq 1/\lambda < \infty$. Therefore:

$$\sigma(e_t : t \in I) \subseteq \sigma(\varepsilon_t : t \in I)$$

for any index set $I \subseteq \mathbb{Z}$. Since α-mixing coefficients are defined as suprema over sub-σ-algebras, the containment gives $\alpha_k(\{e_t\}) \leq \alpha_k(\{\varepsilon_t\})$ for all $k \geq 0$. □

B.3.3 ECDF Concentration under Mixing

Lemma B.17 (Empirical CDF Concentration under Mixing)**.** *Under Assumption B.14, for the true nonconformity scores $\{e_t\}_{t=1}^T$, with probability at least $1 - (M(\log T)^2/(2T))^{1/3}$:*

$$\sup_{x\in\mathbb{R}} \left|\widetilde{F}_{T+1}(x) - F_e(x)\right| \leq \frac{(M/2)^{1/3}(\log T)^{2/3}}{T^{1/3}}. \tag{33}$$

Proof. By Lemma B.16, the scores $\{e_t\}_{t=1}^T$ are stationary and α-mixing with $\sum_{k=0}^{\infty} \alpha_k \leq M$. The result follows from Rio's inequality [27, Proposition 7.1] applied to the indicator processes $\{\mathbb{1}_{\{e_t \leq x\}}\}_{t=1}^T$, following the argument of Xu et al. [40, Lemma B.11]. □

B.3.4 Test Point Error under Known Covariance

Lemma B.18 (Test Point Nonconformity Score Error under Known Mean and Covariance)**.** *Under Assumptions B.14, B.15, B.2, and B.3:*

$$|e_{T+1} - \hat{e}_{T+1}| \leq \tilde{\zeta}_T = \frac{4B\delta_T}{\lambda}. \tag{34}$$

Proof. Under Assumption B.15, $\hat{\mu} = \mu$ and $\widehat{\Sigma} = \Sigma$. In the telescoping decomposition, $T_2^{(T+1)} = 0$ (no mean error) and $T_3^{(T+1)} = 0$ (no covariance error). Only the residual error remains:

$$|e_{T+1} - \hat{e}_{T+1}| = T_1^{(T+1)} \leq \frac{4B}{\lambda} \|\phi(\varepsilon_{T+1}) - \phi(\hat{\varepsilon}_{T+1})\| \leq \frac{4B\delta_T}{\lambda}. \qquad \square$$

B.3.5 Coverage under Mixing

Theorem B.19 (Marginal Coverage Guarantee under α-Mixing)**.** *Under Assumptions B.14, B.15, B.2, and B.3, for any $T \geq 2$, $\alpha \in (0,1)$, and $\lambda > 0$, with probability at least $1 - (M(\log T)^2/(2T))^{1/3}$:*

$$\left|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha}) - (1-\alpha)\right| \leq 16 \cdot \frac{(M/2)^{1/3}(\log T)^{2/3}}{T^{1/3}} + 4(L+1)(\tilde{C}_S + \tilde{\zeta}_T),$$

where $\tilde{C}_S = \sqrt{4B\delta_T/\lambda}$ and $\tilde{\zeta}_T = 4B\delta_T/\lambda = \tilde{C}_S^2$.

Under α-mixing dependence, X_{T+1} typically contains lags of Y_t and is therefore correlated with ε_{T+1}. The exogeneity condition $\varepsilon_{T+1} \perp X_{T+1}$ required for conditional coverage (Assumption B.1) fails in this setting, and the bound is strictly marginal. Exact conditional coverage is known to be unachievable distribution-free [38, 11]. Approximate conditional coverage is assessed empirically via worst-slab coverage in Section 5.1.

Proof. Under Assumption B.15, the same argument as Lemma B.18 applies to each calibration point: for $t = 1, \ldots, T$,

$$|e_t - \hat{e}_t| \leq \frac{4B\delta_T}{\lambda} = \tilde{C}_S^2,$$

so $(1/T)\sum_{t=1}^T |e_t - \hat{e}_t| \leq \tilde{C}_S^2$.

By the same argument as Lemma B.10 (which does not rely on independence of the sequence):

$$\sup_x |\widehat{F}_{T+1}(x) - \widetilde{F}_{T+1}(x)| \leq (L+1)\tilde{C}_S + 2\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)|.$$

Define $A_T = \{\sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| \leq (M/2)^{1/3}(\log T)^{2/3}/T^{1/3}\}$. By Lemma B.17:

$$\mathbb{P}(A_T) \geq 1 - \left(\frac{M(\log T)^2}{2T}\right)^{1/3}.$$

Following the framework of Theorem B.12, conditioning on A_T:

$$\begin{aligned}
|\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})| &\leq \sup_x |\widehat{F}_{T+1}(x) - \widetilde{F}_{T+1}(x)| + \sup_x |\widetilde{F}_{T+1}(x) - F_e(x)| + L\tilde{\zeta}_T \\
&\leq (L+1)\tilde{C}_S + 3 \cdot \frac{(M/2)^{1/3}(\log T)^{2/3}}{T^{1/3}} + L\tilde{\zeta}_T \\
&\leq (L+1)(\tilde{C}_S + \tilde{\zeta}_T) + 3 \cdot \frac{(M/2)^{1/3}(\log T)^{2/3}}{T^{1/3}}.
\end{aligned}$$

Since $F_e(e_{T+1}) \sim \text{Uniform}(0,1)$ under stationarity, $\mathbb{P}(|F_e(e_{T+1}) - \gamma| \le \eta \mid A_T) \le 2\eta$. Note also that

$$\mathbb{P}(A_T^c) \le \left(\frac{M(\log T)^2}{2T}\right)^{1/3} < 2 \cdot \frac{(M/2)^{1/3}(\log T)^{2/3}}{T^{1/3}} \qquad \text{for } T \ge 2.$$

Therefore, for any $\gamma \in [0,1]$:

$$\begin{aligned}&\mathbb{P}\big(|F_e(e_{T+1}) - \gamma| \le |\widehat{F}_{T+1}(\hat{e}_{T+1}) - F_e(e_{T+1})|\big)\\ &\le 2(L+1)(\tilde{C}_S + \tilde{\zeta}_T) + 8 \cdot \frac{(M/2)^{1/3}(\log T)^{2/3}}{T^{1/3}}.\end{aligned}$$

Applying with $\gamma = \beta$ and $\gamma = 1 - \alpha + \beta$ and summing gives the stated bound. □

B.3.6 Convergence Rate under Mixing

Theorem B.20 (Marginal Coverage Convergence Rate under α-Mixing)**.** *Under Assumptions B.14, B.15, B.2, and B.3, there exists $C > 0$ depending only on B, M, L such that with probability at least $1 - (M(\log T)^2/(2T))^{1/3}$:*

$$\left|\mathbb{P}(Y_{T+1} \in \widehat{C}^{\alpha}_{T+1}) - (1-\alpha)\right| \le \begin{cases} C \cdot T^{-1/3}(\log T)^{2/3} & \text{if } \rho > 2/3,\ \lambda = T^{-\beta},\ \beta \in (0, \rho - 2/3) \\ C \cdot T^{(\beta-\rho)/2} & \text{if } \rho \le 2/3,\ \lambda = T^{-\beta},\ \beta \in (0, \rho) \end{cases}$$

Proof. With $\delta_T = O(T^{-\rho})$ and $\lambda = T^{-\beta}$:

$$\tilde{C}_S = \sqrt{\frac{4B\delta_T}{\lambda}} = O(T^{(\beta-\rho)/2}), \qquad \tilde{\zeta}_T = \frac{4B\delta_T}{\lambda} = O(T^{\beta-\rho}) \tag{35}$$

Since $\beta < \rho$, both vanish as $T \to \infty$. For large T where $\tilde{C}_S < 1$:

$$\tilde{C}_S + \tilde{\zeta}_T = \tilde{C}_S(1 + \tilde{C}_S) \le 2\tilde{C}_S$$

so $4(L+1)(\tilde{C}_S + \tilde{\zeta}_T) = O(T^{(\beta-\rho)/2})$.

From Theorem B.19:

$$\left|\mathbb{P}(Y_{T+1} \in \widehat{C}^{\alpha}_{T+1}) - (1-\alpha)\right| \le \underbrace{O(T^{-1/3}(\log T)^{2/3})}_{\text{mixing ECDF term}} + \underbrace{O(T^{(\beta-\rho)/2})}_{\text{score approximation term}}.$$

The score term decays faster than the mixing term when:

$$\frac{\beta - \rho}{2} < -\frac{1}{3} \quad \Longleftrightarrow \quad \beta < \rho - \frac{2}{3}.$$

Case $\rho > 2/3$. We can choose $\beta \in (0, \rho - 2/3)$. Then:

$$\frac{\beta - \rho}{2} < \frac{(\rho - 2/3) - \rho}{2} = -\frac{1}{3},$$

so the mixing ECDF term dominates, giving rate $O(T^{-1/3}(\log T)^{2/3})$.

Case $\rho \le 2/3$. We choose $\beta \in (0, \rho)$. Since $\rho \le 2/3$, for any $\beta > 0$:

$$\frac{\beta - \rho}{2} \ge \frac{-\rho}{2} \ge -\frac{1}{3},$$

so the score term dominates, giving rate $O(T^{(\beta-\rho)/2})$. □

C Evaluation Details and Additional Experimental Results

C.1 Volume Estimation

For the Mahalanobis score, the prediction region is an ellipsoid and its volume admits a closed-form expression

$$\mathrm{V} = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \cdot \hat{q}^{d/2} \cdot \det(\widehat{\Sigma})^{1/2}$$

where $\hat{q}$ is the conformal threshold and $\widehat{\Sigma}$ is the sample covariance. For all other methods, the prediction region has no closed form and we estimate its volume by Monte Carlo. We sample uniformly from a bounding box constructed by padding the range of the calibration residuals by 1% in each coordinate, score each sample, and compute the fraction falling inside the prediction region. We use $50{,}000$ MC samples. We use the same random samples (controlled by a fixed seed) for all scorers within each experimental seed. In the real data experiments, we instead sample from an enlarged Mahalanobis ellipsoid.

C.2 Worst-Slab Coverage

Worst-slab coverage (WSC) [7] approximates conditional coverage by searching for subgroups of the test data with low empirical coverage. Let $c_i = \mathbf{1}\{Y_i \in C(X_i)\}$ denote the coverage indicator for test point i. We consider slabs of the form $\{x : a \le v^\top x \le b\}$ where v is a unit direction. For each direction v, we project the test features onto v and compute the empirical quantiles of the projections $\{v^\top X_i\}$. We construct slabs that contain approximately 20% of the data. Worst-slab coverage is then defined as

$$\mathrm{WSC} = \min_{v,q} \frac{\sum_i c_i \, \mathbf{1}\{Q_v(q) \le v^\top X_i \le Q_v(q+0.2)\}}{\sum_i \mathbf{1}\{Q_v(q) \le v^\top X_i \le Q_v(q+0.2)\}}.$$

In practice, we approximate this search by sampling 200 random unit directions and evaluating 20 quantile positions per direction. Each slab is required to contain at least 30 test points. The same set of random directions is used for all methods and random seeds.

C.3 Additional Details and Results on Synthetic Data

C.3.1 Data Generating Process

We consider a bivariate regression task with a single input $X \sim \mathrm{Uniform}(0,1)$ and response $Y = f(X) + \varepsilon \in \mathbb{R}^2$, where the true regression function is

$$f_1(x) = 3x^2 - 1.5x + \sin(4\pi x),$$
$$f_2(x) = 2x^3 - x + 0.5\cos(3\pi x).$$

The noise is constructed as $\varepsilon = (z_1,\ z_2 + 0.5\, z_1^2 - 0.5\sigma_1^2)$ where $z_1 \sim \mathcal{N}(0, 0.3^2)$ and $z_2 \sim \mathcal{N}(0, 0.1^2)$ are independent, and $\sigma_1 = 0.3$. The noise is also centered by subtracting its sample mean. We generate $n = 10{,}000$ samples and use a 50/25/25 train/calibration/test split.

C.3.2 Models

We use two fitted models to compare the conformal prediction results: a simple ordinary least squares (`LinearRegression`) and a more complex fitting via a two-layer MLP with hidden sizes $(64, 64)$ trained for up to 1000 iterations.

C.4 Additional Details and Results on Real Data

C.4.1 Datasets

We use three datasets previously used in Feldman et al. [10] and Romano et al. [28]: House (King County house sales, Kaggle) [14], Bio (physicochemical properties of protein tertiary structure, UCI #265) [25], and Blog (BlogFeedback, UCI #304) [6]. Following Feldman et al. [10], we augment the original univariate response with one additional variable to obtain a bivariate target ($d = 2$). We

further extend each dataset to $d = 3$ and $d = 4$ by including additional response variables from the same source, listed in Table 3. On House we also consider $d = 6, 8$ and 10, described below.

Dataset	n	Response	Additional response 1	Add. response 2	Add. response 3
House	21,613	price	lat	sqft_living15	sqft_lot15
Bio	45,730	RMSD	F7	F8	F9
Blog	52,397	col. 281	col. 61	col. 62	col. 280

Table 3: Dataset characteristics and target variables.

The House settings with $d = 6, 8$ and 10 use nested sets of continuous response variables, listed in Table 4. The variable sqft_above is removed, because sqft_living = sqft_above + sqft_basement holds exactly and would make the residual covariance singular.

d	Response variables
6	price, lat, long, sqft_living, sqft_living15, sqft_lot
8	those at $d = 6$, sqft_lot15, grade
10	those at $d = 8$, bathrooms, bedrooms

Table 4: Response variables for House at $d = 6, 8, 10$.

All datasets are split $50/25/25$ into training, calibration, and test sets, and the calibration set is further divided into two equal folds: a fit fold, on which the score is constructed, and a quantile fold, on which the threshold is computed. Results are averaged over 50 random seeds controlling the split permutation. Full results including the comparison among all methods, coverages, volume and WSC are provided in Section C.4.3.

C.4.2 Models

We also use two models for our real data experiment, with a simpler and a more complex model choice. Ridge regression (regularization parameter chosen to be 1.0) and a two-layer MLP with hidden sizes $(128, 64)$ trained for up to 500 iterations. For multivariate targets, Ridge is wrapped in `MultiOutputRegressor`, while the MLP handles multi-output natively. All models are implemented with SciKit-Learn. The two models represent different levels of misspecification: Ridge regression cannot capture the nonlinear structure in the data and produces residuals with pronounced non-elliptical geometry. This allows us to evaluate the MKS under different fitting regimes.

C.4.3 Full Results

In Tables 5–16, we report coverage, volume, and worst-slab-coverage (WSC) for all datasets, dimensions, coverage levels, and models as described in Appendix C.4.1. Under Ridge, the MKS achieves the smallest volumes across all configurations. Under the MLP it does so in all but seven configurations, all at $\alpha \leq 0.02$ and six of them on Blog, where the density-only score is smaller by up to 12%. The Bonferroni method produces the largest regions in all cases. The density-only score and Mahalanobis ellipsoid are closer in performance, with the relative ordering depending on the dataset and dimension. Coverage is maintained at the nominal level across all methods, consistent with the distribution-free guarantee. WSC values are comparable across methods at the tighter coverage levels. At $\alpha = 0.1$ and 0.05 the MKS has the lowest WSC in most configurations, consistent with the non-linear, anisotropic nature of the score as mentioned in Section 5.1.

Bio $d = 2$

		Ridge			MLP		
α	Method	Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	0.90024±0.00442	6.311±2.211	0.80853±0.01461	0.89990±0.00378	5.612±2.048	0.84163±0.01323
	Mahalanobis	0.89963±0.00468	5.258±1.704	0.79697±0.01796	0.89997±0.00390	4.621±1.594	0.83398±0.01552
	Density (med/2)	0.89975±0.00500	3.461±0.191	0.77721±0.01251	0.89962±0.00446	2.773±0.114	0.81109±0.00927
	OT-CP	0.90047±0.00542	3.277±0.218	0.77165±0.01183	0.89943±0.00468	2.612±0.134	0.80339±0.01055
	MKS (med/2)	0.89959±0.00639	**2.706±0.217**	0.72916±0.01516	0.90004±0.00481	**2.094±0.138**	0.77356±0.01256
0.05	Bonferroni	0.95035±0.00363	8.954±3.127	0.88657±0.01092	0.95047±0.00282	8.639±3.111	0.90546±0.01137
	Mahalanobis	0.95069±0.00350	7.501±2.367	0.88007±0.01245	0.95029±0.00282	7.091±2.419	0.89989±0.01252
	Density (med/2)	0.95037±0.00351	5.002±0.294	0.86471±0.00955	0.95035±0.00359	4.208±0.233	0.87950±0.00755
	OT-CP	0.95037±0.00364	4.747±0.313	0.86051±0.00939	0.95019±0.00355	4.194±0.256	0.87844±0.00783
	MKS (med/2)	0.95005±0.00416	**4.048±0.356**	0.84019±0.01118	0.94959±0.00387	**3.382±0.263**	0.85843±0.01054
0.02	Bonferroni	0.98011±0.00221	13.48±4.43	0.94049±0.00686	0.97957±0.00263	13.28±4.61	0.94918±0.00926
	Mahalanobis	0.97998±0.00239	11.23±3.42	0.93718±0.00800	0.97967±0.00251	10.94±3.56	0.94702±0.00940
	Density (med/2)	0.97982±0.00236	7.510±0.538	0.93097±0.00657	0.97996±0.00259	6.715±0.508	0.93433±0.00618
	OT-CP	0.97990±0.00229	7.433±0.622	0.92993±0.00659	0.97994±0.00258	7.066±0.545	0.93649±0.00631
	MKS (med/2)	0.97995±0.00234	**6.582±0.636**	0.92268±0.00784	0.97986±0.00275	**6.137±0.678**	0.92831±0.00830
0.01	Bonferroni	0.98988±0.00188	18.79±5.73	0.96245±0.00635	0.98990±0.00164	18.05±5.52	0.96848±0.00691
	Mahalanobis	0.98994±0.00193	15.72±4.47	0.96154±0.00650	0.98990±0.00167	14.94±4.35	0.96714±0.00715
	Density (med/2)	0.98994±0.00174	11.63±1.64	0.95798±0.00585	0.98996±0.00183	**10.58±1.32**	0.96003±0.00602
	OT-CP	0.99010±0.00165	12.03±1.71	0.95864±0.00577	0.98985±0.00176	10.89±1.21	0.96087±0.00543
	MKS (med/2)	0.98993±0.00167	**11.16±1.53**	0.95697±0.00672	0.98990±0.00196	10.58±1.52	0.95963±0.00687

Table 5: Bio dataset ($d = 2$), Ridge and MLP. Mean ± std over 50 seeds.

Bio $d = 3$

		Ridge			MLP		
α	Method	Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	0.90062±0.00514	28.30±9.91	0.67673±0.01274	0.90014±0.00559	20.45±7.87	0.76367±0.01208
	Mahalanobis	0.90055±0.00541	20.55±6.65	0.68561±0.01514	0.90052±0.00541	14.86±5.34	0.76042±0.01385
	Density (med/2)	0.90039±0.00550	15.55±0.70	0.68509±0.01414	0.90055±0.00533	10.50±0.62	0.75755±0.01071
	OT-CP	0.90061±0.00539	13.48±0.75	0.68072±0.01380	0.90007±0.00568	9.187±0.649	0.74698±0.01133
	MKS (med/2)	0.89971±0.00564	**9.164±0.689**	0.66672±0.01675	0.90030±0.00523	**6.266±0.542**	0.72609±0.01280
0.05	Bonferroni	0.95011±0.00364	47.19±16.40	0.82223±0.01145	0.95053±0.00311	36.90±13.44	0.85860±0.00850
	Mahalanobis	0.95021±0.00405	32.81±10.29	0.82354±0.01321	0.95029±0.00328	26.08±8.87	0.85698±0.00959
	Density (med/2)	0.95024±0.00404	24.49±1.21	0.82114±0.01191	0.95053±0.00378	18.03±1.06	0.85571±0.00947
	OT-CP	0.94998±0.00430	22.07±1.29	0.81398±0.01450	0.95048±0.00378	16.52±1.12	0.85166±0.00967
	MKS (med/2)	0.95023±0.00383	**14.54±1.12**	0.81516±0.01418	0.95044±0.00418	**11.54±1.24**	0.83632±0.01182
0.02	Bonferroni	0.97997±0.00265	94.48±28.83	0.91431±0.01070	0.97990±0.00223	67.81±22.45	0.92898±0.00755
	Mahalanobis	0.98014±0.00272	62.88±18.33	0.91495±0.01145	0.98013±0.00234	46.44±14.71	0.92951±0.00805
	Density (med/2)	0.98017±0.00283	44.66±2.24	0.91558±0.01147	0.97985±0.00242	30.94±1.98	0.92737±0.00753
	OT-CP	0.97999±0.00242	34.79±2.18	0.91913±0.00896	0.97997±0.00228	31.30±2.79	0.92691±0.00739
	MKS (med/2)	0.97998±0.00227	**26.06±2.63**	0.91575±0.00923	0.98001±0.00266	**23.83±2.80**	0.92131±0.00922
0.01	Bonferroni	0.98998±0.00172	156.9±47.5	0.95456±0.00740	0.98993±0.00181	102.4±30.2	0.95970±0.00658
	Mahalanobis	0.99004±0.00166	94.07±28.47	0.95569±0.00723	0.98992±0.00171	68.50±19.14	0.95964±0.00624
	Density (med/2)	0.99003±0.00171	59.25±5.96	0.95578±0.00705	0.98988±0.00186	47.56±5.66	0.95816±0.00683
	OT-CP	0.99004±0.00147	57.10±8.66	0.95597±0.00641	0.98990±0.00182	52.08±7.32	0.95836±0.00704
	MKS (med/2)	0.99014±0.00175	**45.42±6.14**	0.95629±0.00810	0.98994±0.00160	**41.76±6.07**	0.95736±0.00656

Table 6: Bio dataset ($d = 3$), Ridge and MLP. Mean ± std over 50 seeds.

Bio $d = 4$

		Ridge			MLP		
α	Method	Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	0.90077±0.00526	65.18±22.63	0.69891±0.01348	0.90028±0.00559	46.96±16.59	0.76188±0.01405
	Mahalanobis	0.89917±0.00567	38.20±11.95	0.70354±0.01472	0.89971±0.00571	28.22±9.44	0.75492±0.01611
	Density (med/2)	0.89970±0.00534	34.91±1.55	0.69797±0.01306	0.89980±0.00582	24.23±1.51	0.75685±0.01399
	OT-CP	0.90060±0.00521	28.66±2.01	0.68825±0.01320	0.90014±0.00539	20.85±1.55	0.74028±0.01362
	MKS (med/2)	0.89978±0.00618	**14.47±1.21**	0.70254±0.01783	0.90004±0.00520	**11.31±0.96**	0.72206±0.01351
0.05	Bonferroni	0.95007±0.00388	128.5±41.8	0.84640±0.01118	0.95002±0.00338	106.7±35.2	0.87083±0.00925
	Mahalanobis	0.95047±0.00393	76.97±23.65	0.84213±0.01232	0.95061±0.00346	58.52±18.71	0.86828±0.01035
	Density (med/2)	0.95054±0.00399	65.15±3.98	0.83334±0.01190	0.95077±0.00369	47.45±3.00	0.86612±0.00867
	OT-CP	0.95041±0.00403	57.97±3.92	0.83158±0.01309	0.95005±0.00370	45.25±3.83	0.85523±0.00987
	MKS (med/2)	0.95023±0.00400	**27.68±2.72**	0.84483±0.01222	0.95017±0.00434	**24.86±2.68**	0.84782±0.01251
0.02	Bonferroni	0.98008±0.00243	404.9±121.7	0.93363±0.00864	0.97998±0.00263	257.3±74.2	0.94541±0.00830
	Mahalanobis	0.98032±0.00235	176.6±54.9	0.93530±0.00824	0.98007±0.00218	130.0±35.8	0.94427±0.00786
	Density (med/2)	0.98016±0.00248	130.5±6.8	0.92616±0.00891	0.98010±0.00239	100.7±10.7	0.93666±0.00715
	OT-CP	0.98012±0.00219	138.7±17.8	0.92801±0.00758	0.97995±0.00262	136.8±22.8	0.93322±0.00812
	MKS (med/2)	0.98017±0.00250	**69.39±8.85**	0.92747±0.00810	0.97985±0.00237	**64.46±8.26**	0.92845±0.00797
0.01	Bonferroni	0.98986±0.00160	903.4±284.3	0.96764±0.00577	0.98996±0.00149	713.3±223.2	0.97199±0.00529
	Mahalanobis	0.98993±0.00148	423.7±115.7	0.96591±0.00557	0.99006±0.00143	323.4±77.4	0.97107±0.00548
	Density (med/2)	0.98989±0.00172	277.7±64.5	0.95959±0.00608	0.99004±0.00155	212.1±33.7	0.96283±0.00499
	OT-CP	0.99004±0.00147	392.4±97.8	0.96110±0.00532	0.98993±0.00154	385.8±102.0	0.96414±0.00506
	MKS (med/2)	0.99004±0.00164	**140.6±24.3**	0.96004±0.00605	0.98996±0.00163	**133.9±24.2**	0.96087±0.00565

Table 7: Bio dataset ($d = 4$), Ridge and MLP. Mean ± std over 50 seeds.

House $d = 2$

		Ridge			MLP		
α	Method	Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	0.90006±0.00668	6.937±0.293	0.77850±0.01488	0.89912±0.00650	4.788±0.217	0.77956±0.01492
	Mahalanobis	0.89887±0.00633	6.084±0.248	0.75310±0.01587	0.89935±0.00720	4.198±0.189	0.76938±0.01503
	Density (med/2)	0.89988±0.00704	5.856±0.164	0.75724±0.01467	0.89885±0.00697	4.291±0.199	0.77813±0.01507
	OT-CP	0.90015±0.00651	5.831±0.165	0.76022±0.01329	0.89897±0.00702	4.219±0.202	0.77495±0.01534
	MKS (med/2)	0.90038±0.00632	**5.314±0.175**	0.73602±0.01514	0.89945±0.00764	**3.897±0.188**	0.75428±0.01582
0.05	Bonferroni	0.95039±0.00420	9.621±0.406	0.85574±0.01102	0.94965±0.00488	7.606±0.403	0.86303±0.01333
	Mahalanobis	0.95004±0.00464	8.532±0.381	0.84627±0.01257	0.95021±0.00540	6.596±0.367	0.85656±0.01439
	Density (med/2)	0.94993±0.00434	8.314±0.308	0.85012±0.01226	0.94977±0.00534	6.630±0.332	0.86250±0.01426
	OT-CP	0.95020±0.00434	8.248±0.303	0.85312±0.01233	0.94987±0.00529	6.676±0.357	0.86309±0.01421
	MKS (med/2)	0.95031±0.00436	**7.721±0.273**	0.83944±0.01378	0.95001±0.00531	**6.183±0.321**	0.84873±0.01557
0.02	Bonferroni	0.97973±0.00302	15.74±1.60	0.91471±0.01016	0.97996±0.00324	13.27±1.00	0.92823±0.01071
	Mahalanobis	0.97984±0.00298	13.32±1.31	0.91243±0.01101	0.98008±0.00305	11.03±0.74	0.92474±0.01115
	Density (med/2)	0.98002±0.00293	12.68±0.95	0.91521±0.01018	0.98021±0.00339	10.68±0.73	0.92903±0.01118
	OT-CP	0.97994±0.00256	12.06±0.66	0.91772±0.00821	0.98022±0.00319	10.93±0.77	0.93110±0.01080
	MKS (med/2)	0.97987±0.00278	**11.19±0.64**	0.91556±0.01001	0.98036±0.00317	**10.14±0.80**	0.92551±0.01266
0.01	Bonferroni	0.98979±0.00245	27.45±5.06	0.95143±0.01046	0.98994±0.00249	19.73±2.44	0.95680±0.00975
	Mahalanobis	0.98973±0.00263	22.39±3.98	0.95079±0.01138	0.99006±0.00211	16.55±2.01	0.95602±0.00898
	Density (med/2)	0.98989±0.00250	18.00±1.82	0.95169±0.01051	0.99030±0.00220	14.59±1.30	0.95859±0.00883
	OT-CP	0.98997±0.00252	17.31±2.15	0.95225±0.01074	0.99039±0.00230	15.15±1.54	0.96004±0.00884
	MKS (med/2)	0.98999±0.00229	**15.41±1.43**	0.95345±0.00941	0.99026±0.00219	**14.24±1.48**	0.95861±0.00912

Table 8: House dataset ($d = 2$), Ridge and MLP. Mean ± std over 50 seeds.

House $d = 3$

α	Method	Ridge			MLP		
		Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	$0.90065_{\pm 0.00657}$	$20.85_{\pm 1.16}$	$0.73362_{\pm 0.01642}$	$0.90169_{\pm 0.00744}$	$14.93_{\pm 0.89}$	$0.75628_{\pm 0.01503}$
	Mahalanobis	$0.90021_{\pm 0.00731}$	$16.75_{\pm 0.83}$	$0.72011_{\pm 0.01762}$	$0.90104_{\pm 0.00726}$	$11.21_{\pm 0.69}$	$0.73932_{\pm 0.01643}$
	Density (med/2)	$0.89993_{\pm 0.00691}$	$16.79_{\pm 0.68}$	$0.73210_{\pm 0.01571}$	$0.90020_{\pm 0.00691}$	$11.97_{\pm 0.68}$	$0.75362_{\pm 0.01476}$
	OT-CP	$0.90088_{\pm 0.00682}$	$17.08_{\pm 0.80}$	$0.73737_{\pm 0.01567}$	$0.90081_{\pm 0.00701}$	$11.67_{\pm 0.69}$	$0.74997_{\pm 0.01453}$
	MKS (med/2)	$0.90041_{\pm 0.00700}$	$\mathbf{14.79}_{\pm \mathbf{0.70}}$	$0.71360_{\pm 0.01933}$	$0.90137_{\pm 0.00722}$	$\mathbf{10.32}_{\pm \mathbf{0.68}}$	$0.72964_{\pm 0.01702}$
0.05	Bonferroni	$0.95053_{\pm 0.00502}$	$38.09_{\pm 2.92}$	$0.83451_{\pm 0.01292}$	$0.95058_{\pm 0.00527}$	$28.93_{\pm 2.40}$	$0.85234_{\pm 0.01524}$
	Mahalanobis	$0.94997_{\pm 0.00462}$	$28.18_{\pm 2.15}$	$0.82355_{\pm 0.01408}$	$0.95024_{\pm 0.00474}$	$20.66_{\pm 1.47}$	$0.83902_{\pm 0.01542}$
	Density (med/2)	$0.95030_{\pm 0.00472}$	$28.30_{\pm 1.96}$	$0.83277_{\pm 0.01244}$	$0.95050_{\pm 0.00511}$	$21.46_{\pm 1.61}$	$0.84982_{\pm 0.01451}$
	OT-CP	$0.95047_{\pm 0.00487}$	$27.76_{\pm 1.91}$	$0.83325_{\pm 0.01289}$	$0.95050_{\pm 0.00551}$	$21.07_{\pm 1.75}$	$0.84755_{\pm 0.01519}$
	MKS (med/2)	$0.95067_{\pm 0.00448}$	$\mathbf{25.67}_{\pm \mathbf{1.69}}$	$0.82716_{\pm 0.01358}$	$0.95005_{\pm 0.00481}$	$\mathbf{19.11}_{\pm \mathbf{1.66}}$	$0.83608_{\pm 0.01444}$
0.02	Bonferroni	$0.97989_{\pm 0.00333}$	$89.01_{\pm 12.70}$	$0.91889_{\pm 0.01257}$	$0.98036_{\pm 0.00325}$	$62.77_{\pm 6.74}$	$0.92740_{\pm 0.01078}$
	Mahalanobis	$0.97967_{\pm 0.00300}$	$65.02_{\pm 8.24}$	$0.91541_{\pm 0.01130}$	$0.97989_{\pm 0.00363}$	$43.42_{\pm 5.12}$	$0.92144_{\pm 0.01243}$
	Density (med/2)	$0.97956_{\pm 0.00328}$	$57.10_{\pm 6.35}$	$0.91515_{\pm 0.01223}$	$0.98004_{\pm 0.00351}$	$40.72_{\pm 4.32}$	$0.92577_{\pm 0.01155}$
	OT-CP	$0.98028_{\pm 0.00280}$	$55.02_{\pm 5.90}$	$0.91700_{\pm 0.01091}$	$0.98058_{\pm 0.00360}$	$43.22_{\pm 5.03}$	$0.92727_{\pm 0.01198}$
	MKS (med/2)	$0.97996_{\pm 0.00277}$	$\mathbf{48.01}_{\pm \mathbf{4.22}}$	$0.91654_{\pm 0.01075}$	$0.98034_{\pm 0.00331}$	$\mathbf{39.15}_{\pm \mathbf{4.06}}$	$0.92459_{\pm 0.01192}$
0.01	Bonferroni	$0.99025_{\pm 0.00228}$	$177.3_{\pm 33.1}$	$0.95639_{\pm 0.00930}$	$0.99027_{\pm 0.00239}$	$111.5_{\pm 17.1}$	$0.95876_{\pm 0.00897}$
	Mahalanobis	$0.99006_{\pm 0.00225}$	$122.9_{\pm 22.6}$	$0.95425_{\pm 0.00998}$	$0.99024_{\pm 0.00251}$	$76.90_{\pm 10.87}$	$0.95706_{\pm 0.00984}$
	Density (med/2)	$0.99025_{\pm 0.00213}$	$95.54_{\pm 14.11}$	$0.95552_{\pm 0.00927}$	$0.99047_{\pm 0.00265}$	$65.65_{\pm 8.40}$	$0.95919_{\pm 0.00960}$
	OT-CP	$0.99003_{\pm 0.00201}$	$95.13_{\pm 13.46}$	$0.95469_{\pm 0.00907}$	$0.99037_{\pm 0.00241}$	$73.51_{\pm 10.78}$	$0.95922_{\pm 0.00915}$
	MKS (med/2)	$0.99043_{\pm 0.00202}$	$\mathbf{76.75}_{\pm \mathbf{8.92}}$	$0.95730_{\pm 0.00866}$	$0.99014_{\pm 0.00255}$	$\mathbf{65.29}_{\pm \mathbf{9.94}}$	$0.95789_{\pm 0.00941}$

Table 9: House dataset ($d = 3$), Ridge and MLP. Mean $\pm$ std over 50 seeds.

House $d = 4$

α	Method	Ridge			MLP		
		Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	$0.90121_{\pm 0.00770}$	$71.19_{\pm 12.50}$	$0.69704_{\pm 0.02018}$	$0.90116_{\pm 0.00761}$	$53.00_{\pm 7.85}$	$0.71708_{\pm 0.01802}$
	Mahalanobis	$0.90012_{\pm 0.00785}$	$40.15_{\pm 6.51}$	$0.67636_{\pm 0.02025}$	$0.90088_{\pm 0.00793}$	$28.44_{\pm 4.16}$	$0.69745_{\pm 0.01927}$
	Density (med/2)	$0.90085_{\pm 0.00803}$	$33.34_{\pm 2.90}$	$0.68651_{\pm 0.01932}$	$0.90112_{\pm 0.00860}$	$24.81_{\pm 2.49}$	$0.70875_{\pm 0.01903}$
	OT-CP	$0.90028_{\pm 0.00742}$	$35.53_{\pm 3.19}$	$0.69105_{\pm 0.01780}$	$0.90085_{\pm 0.00817}$	$25.66_{\pm 2.79}$	$0.70857_{\pm 0.01814}$
	MKS (med/2)	$0.90203_{\pm 0.00793}$	$\mathbf{18.94}_{\pm \mathbf{2.18}}$	$0.66531_{\pm 0.02129}$	$0.90166_{\pm 0.00689}$	$\mathbf{15.32}_{\pm \mathbf{1.72}}$	$0.67872_{\pm 0.02019}$
0.05	Bonferroni	$0.95047_{\pm 0.00579}$	$201.5_{\pm 45.5}$	$0.80905_{\pm 0.02005}$	$0.95021_{\pm 0.00520}$	$144.3_{\pm 25.0}$	$0.82313_{\pm 0.01833}$
	Mahalanobis	$0.95043_{\pm 0.00578}$	$108.8_{\pm 23.9}$	$0.80305_{\pm 0.02040}$	$0.94960_{\pm 0.00555}$	$74.68_{\pm 12.73}$	$0.81042_{\pm 0.01914}$
	Density (med/2)	$0.95074_{\pm 0.00543}$	$90.74_{\pm 16.63}$	$0.80372_{\pm 0.01746}$	$0.95030_{\pm 0.00606}$	$63.80_{\pm 9.18}$	$0.81654_{\pm 0.01882}$
	OT-CP	$0.95074_{\pm 0.00577}$	$93.90_{\pm 16.02}$	$0.81038_{\pm 0.01808}$	$0.95024_{\pm 0.00573}$	$71.54_{\pm 11.34}$	$0.81983_{\pm 0.01869}$
	MKS (med/2)	$0.95054_{\pm 0.00540}$	$\mathbf{54.95}_{\pm \mathbf{9.18}}$	$0.80144_{\pm 0.01814}$	$0.95076_{\pm 0.00603}$	$\mathbf{48.74}_{\pm \mathbf{7.95}}$	$0.81069_{\pm 0.01925}$
0.02	Bonferroni	$0.98057_{\pm 0.00371}$	$1.194_{\pm 0.464 \times 10^3}$	$0.91500_{\pm 0.01504}$	$0.98053_{\pm 0.00347}$	$646.7_{\pm 166.0}$	$0.91747_{\pm 0.01302}$
	Mahalanobis	$0.98054_{\pm 0.00365}$	$550.0_{\pm 179.0}$	$0.91375_{\pm 0.01502}$	$0.98007_{\pm 0.00341}$	$300.2_{\pm 67.1}$	$0.91356_{\pm 0.01314}$
	Density (med/2)	$0.98055_{\pm 0.00360}$	$384.1_{\pm 76.4}$	$0.91375_{\pm 0.01432}$	$0.98034_{\pm 0.00382}$	$233.2_{\pm 47.4}$	$0.91469_{\pm 0.01478}$
	OT-CP	$0.98031_{\pm 0.00376}$	$465.3_{\pm 174.2}$	$0.91317_{\pm 0.01529}$	$0.98012_{\pm 0.00360}$	$343.4_{\pm 112.3}$	$0.91452_{\pm 0.01421}$
	MKS (med/2)	$0.98002_{\pm 0.00383}$	$\mathbf{218.4}_{\pm \mathbf{49.1}}$	$0.91195_{\pm 0.01606}$	$0.97984_{\pm 0.00398}$	$\mathbf{194.8}_{\pm \mathbf{40.5}}$	$0.91382_{\pm 0.01568}$
0.01	Bonferroni	$0.99024_{\pm 0.00217}$	$4.132_{\pm 1.525 \times 10^3}$	$0.95591_{\pm 0.00982}$	$0.99058_{\pm 0.00238}$	$2.218_{\pm 0.814 \times 10^3}$	$0.95728_{\pm 0.01072}$
	Mahalanobis	$0.99028_{\pm 0.00220}$	$1.635_{\pm 0.524 \times 10^3}$	$0.95610_{\pm 0.01001}$	$0.99038_{\pm 0.00238}$	$886.2_{\pm 280.1}$	$0.95628_{\pm 0.01046}$
	Density (med/2)	$0.99043_{\pm 0.00223}$	$673.7_{\pm 120.9}$	$0.95645_{\pm 0.00998}$	$0.99000_{\pm 0.00254}$	$468.5_{\pm 92.0}$	$0.95499_{\pm 0.01084}$
	OT-CP	$0.99029_{\pm 0.00235}$	$1.919_{\pm 1.148 \times 10^3}$	$0.95545_{\pm 0.01051}$	$0.99025_{\pm 0.00222}$	$1.454_{\pm 0.754 \times 10^3}$	$0.95599_{\pm 0.00973}$
	MKS (med/2)	$0.99008_{\pm 0.00268}$	$\mathbf{472.7}_{\pm \mathbf{132.5}}$	$0.95545_{\pm 0.01168}$	$0.98995_{\pm 0.00271}$	$\mathbf{443.6}_{\pm \mathbf{121.0}}$	$0.95515_{\pm 0.01133}$

Table 10: House dataset ($d = 4$), Ridge and MLP. Mean $\pm$ std over 50 seeds.

House $d = 6$

α	Method	Ridge			MLP		
		Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	$0.89973_{\pm 0.00880}$	$4.416_{\pm 0.910 \times 10^3}$	$0.78980_{\pm 0.01780}$	$0.90122_{\pm 0.00787}$	$2.783_{\pm 0.641 \times 10^3}$	$0.79730_{\pm 0.01600}$
	Mahalanobis	$0.90020_{\pm 0.00830}$	$990.2_{\pm 210.9}$	$0.79480_{\pm 0.01620}$	$0.90081_{\pm 0.00775}$	$762.1_{\pm 159.5}$	$0.78644_{\pm 0.01597}$
	Density (med/2)	$0.90028_{\pm 0.00864}$	$1.126_{\pm 0.159 \times 10^3}$	$0.80352_{\pm 0.01575}$	$0.90132_{\pm 0.00767}$	$775.6_{\pm 104.7}$	$0.78758_{\pm 0.01803}$
	OT-CP	$0.90070_{\pm 0.00782}$	$1.306_{\pm 0.148 \times 10^3}$	$0.79594_{\pm 0.01783}$	$0.90172_{\pm 0.00698}$	$962.4_{\pm 125.7}$	$0.77695_{\pm 0.01757}$
	MKS (med/2)	$0.90138_{\pm 0.00790}$	$\mathbf{215.8}_{\pm \mathbf{29.7}}$	$0.78398_{\pm 0.01725}$	$0.90179_{\pm 0.00715}$	$\mathbf{271.6}_{\pm \mathbf{49.2}}$	$0.76608_{\pm 0.01613}$
0.05	Bonferroni	$0.95024_{\pm 0.00584}$	$2.411_{\pm 0.695 \times 10^4}$	$0.89319_{\pm 0.01146}$	$0.95013_{\pm 0.00611}$	$1.593_{\pm 0.470 \times 10^4}$	$0.89397_{\pm 0.01148}$
	Mahalanobis	$0.95008_{\pm 0.00556}$	$4.194_{\pm 0.898 \times 10^3}$	$0.88426_{\pm 0.01229}$	$0.94995_{\pm 0.00553}$	$2.940_{\pm 0.644 \times 10^3}$	$0.88167_{\pm 0.01115}$
	Density (med/2)	$0.95054_{\pm 0.00582}$	$4.285_{\pm 0.867 \times 10^3}$	$0.88855_{\pm 0.01207}$	$0.95077_{\pm 0.00570}$	$2.896_{\pm 0.647 \times 10^3}$	$0.88430_{\pm 0.01196}$
	OT-CP	$0.95064_{\pm 0.00562}$	$4.929_{\pm 1.050 \times 10^3}$	$0.88564_{\pm 0.01240}$	$0.95071_{\pm 0.00484}$	$4.240_{\pm 0.866 \times 10^3}$	$0.87898_{\pm 0.01094}$
	MKS (med/2)	$0.95037_{\pm 0.00524}$	$\mathbf{853.6}_{\pm \mathbf{164.4}}$	$0.88327_{\pm 0.01247}$	$0.95068_{\pm 0.00522}$	$\mathbf{1.115}_{\pm \mathbf{0.246} \times 10^3}$	$0.87543_{\pm 0.01227}$
0.02	Bonferroni	$0.98005_{\pm 0.00340}$	$2.249_{\pm 1.109 \times 10^5}$	$0.94974_{\pm 0.00821}$	$0.98041_{\pm 0.00354}$	$1.207_{\pm 0.474 \times 10^5}$	$0.95114_{\pm 0.00830}$
	Mahalanobis	$0.97993_{\pm 0.00360}$	$2.524_{\pm 0.881 \times 10^4}$	$0.94614_{\pm 0.00900}$	$0.98023_{\pm 0.00351}$	$1.708_{\pm 0.540 \times 10^4}$	$0.94577_{\pm 0.00851}$
	Density (med/2)	$0.98014_{\pm 0.00370}$	$2.832_{\pm 0.675 \times 10^4}$	$0.94988_{\pm 0.00859}$	$0.98030_{\pm 0.00320}$	$1.686_{\pm 0.343 \times 10^4}$	$0.94867_{\pm 0.00748}$
	OT-CP	$0.97987_{\pm 0.00334}$	$5.477_{\pm 2.929 \times 10^4}$	$0.94963_{\pm 0.00788}$	$0.98002_{\pm 0.00349}$	$4.867_{\pm 3.415 \times 10^4}$	$0.94644_{\pm 0.00843}$
	MKS (med/2)	$0.98023_{\pm 0.00392}$	$\mathbf{4.506}_{\pm \mathbf{1.354} \times 10^3}$	$0.94855_{\pm 0.00922}$	$0.98001_{\pm 0.00365}$	$\mathbf{5.563}_{\pm \mathbf{1.635} \times 10^3}$	$0.94447_{\pm 0.00966}$
0.01	Bonferroni	$0.99007_{\pm 0.00252}$	$8.792_{\pm 3.779 \times 10^5}$	$0.97224_{\pm 0.00624}$	$0.99027_{\pm 0.00244}$	$5.119_{\pm 2.356 \times 10^5}$	$0.97216_{\pm 0.00605}$
	Mahalanobis	$0.99010_{\pm 0.00235}$	$9.273_{\pm 3.445 \times 10^4}$	$0.97126_{\pm 0.00549}$	$0.99041_{\pm 0.00244}$	$6.531_{\pm 2.483 \times 10^4}$	$0.97151_{\pm 0.00635}$
	Density (med/2)	$0.99031_{\pm 0.00258}$	$5.411_{\pm 1.174 \times 10^4}$	$0.97232_{\pm 0.00612}$	$0.99028_{\pm 0.00271}$	$3.363_{\pm 0.863 \times 10^4}$	$0.97118_{\pm 0.00640}$
	OT-CP	$0.98963_{\pm 0.00260}$	$3.494_{\pm 2.216 \times 10^5}$	$0.97149_{\pm 0.00584}$	$0.98990_{\pm 0.00236}$	$3.801_{\pm 4.500 \times 10^5}$	$0.97048_{\pm 0.00549}$
	MKS (med/2)	$0.98999_{\pm 0.00267}$	$\mathbf{1.255}_{\pm \mathbf{0.489} \times 10^4}$	$0.97152_{\pm 0.00600}$	$0.99020_{\pm 0.00241}$	$\mathbf{1.604}_{\pm \mathbf{0.557} \times 10^4}$	$0.97045_{\pm 0.00639}$

Table 11: House dataset ($d = 6$), Ridge and MLP. Mean $\pm$ std over 50 seeds.

House $d = 8$

α	Method	Ridge			MLP		
		Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	$0.89947_{\pm 0.00916}$	$1.033_{\pm 0.310 \times 10^5}$	$0.80330_{\pm 0.01771}$	$0.90100_{\pm 0.00850}$	$6.542_{\pm 2.157 \times 10^4}$	$0.80289_{\pm 0.01687}$
	Mahalanobis	$0.90055_{\pm 0.00815}$	$6.866_{\pm 2.303 \times 10^3}$	$0.80401_{\pm 0.01466}$	$0.90152_{\pm 0.00739}$	$5.045_{\pm 1.553 \times 10^3}$	$0.79315_{\pm 0.01462}$
	Density (med/2)	$0.90068_{\pm 0.00868}$	$1.172_{\pm 0.228 \times 10^4}$	$0.80822_{\pm 0.01615}$	$0.90077_{\pm 0.00717}$	$7.121_{\pm 1.355 \times 10^3}$	$0.78692_{\pm 0.01511}$
	OT-CP	$0.89972_{\pm 0.00832}$	$2.215_{\pm 0.415 \times 10^4}$	$0.79501_{\pm 0.01731}$	$0.90134_{\pm 0.00787}$	$1.518_{\pm 0.318 \times 10^4}$	$0.78045_{\pm 0.01779}$
	MKS (med/2)	$0.90165_{\pm 0.00743}$	$\mathbf{895.3}_{\pm \mathbf{165.0}}$	$0.79750_{\pm 0.01543}$	$0.90161_{\pm 0.00809}$	$\mathbf{1.300}_{\pm \mathbf{0.322} \times 10^3}$	$0.77334_{\pm 0.01806}$
0.05	Bonferroni	$0.95048_{\pm 0.00625}$	$1.161_{\pm 0.452 \times 10^6}$	$0.89366_{\pm 0.01335}$	$0.95102_{\pm 0.00573}$	$7.116_{\pm 2.783 \times 10^5}$	$0.89522_{\pm 0.01194}$
	Mahalanobis	$0.95068_{\pm 0.00575}$	$4.231_{\pm 1.497 \times 10^4}$	$0.88761_{\pm 0.01234}$	$0.95098_{\pm 0.00604}$	$3.301_{\pm 1.261 \times 10^4}$	$0.88495_{\pm 0.01180}$
	Density (med/2)	$0.95070_{\pm 0.00639}$	$8.709_{\pm 3.517 \times 10^4}$	$0.89332_{\pm 0.01438}$	$0.95097_{\pm 0.00611}$	$5.513_{\pm 2.040 \times 10^4}$	$0.88989_{\pm 0.01415}$
	OT-CP	$0.95035_{\pm 0.00605}$	$1.739_{\pm 0.699 \times 10^5}$	$0.89365_{\pm 0.01340}$	$0.95097_{\pm 0.00518}$	$1.427_{\pm 0.695 \times 10^5}$	$0.88905_{\pm 0.01277}$
	MKS (med/2)	$0.95123_{\pm 0.00551}$	$\mathbf{7.834}_{\pm \mathbf{2.737} \times 10^3}$	$0.89454_{\pm 0.01243}$	$0.95082_{\pm 0.00567}$	$\mathbf{1.094}_{\pm \mathbf{0.358} \times 10^4}$	$0.88386_{\pm 0.01346}$
0.02	Bonferroni	$0.98047_{\pm 0.00372}$	$5.376_{\pm 4.931 \times 10^7}$	$0.95150_{\pm 0.00857}$	$0.98049_{\pm 0.00376}$	$2.557_{\pm 2.178 \times 10^7}$	$0.95124_{\pm 0.00851}$
	Mahalanobis	$0.98023_{\pm 0.00377}$	$1.170_{\pm 0.810 \times 10^6}$	$0.95001_{\pm 0.00938}$	$0.98042_{\pm 0.00363}$	$7.329_{\pm 4.965 \times 10^5}$	$0.95005_{\pm 0.00916}$
	Density (med/2)	$0.98079_{\pm 0.00364}$	$1.288_{\pm 0.393 \times 10^6}$	$0.95276_{\pm 0.00825}$	$0.98085_{\pm 0.00374}$	$6.954_{\pm 2.036 \times 10^5}$	$0.95214_{\pm 0.00843}$
	OT-CP	$0.98054_{\pm 0.00349}$	$2.571_{\pm 5.353 \times 10^7}$	$0.95234_{\pm 0.00795}$	$0.98036_{\pm 0.00345}$	$1.185_{\pm 2.067 \times 10^7}$	$0.95098_{\pm 0.00859}$
	MKS (med/2)	$0.98041_{\pm 0.00387}$	$\mathbf{1.120}_{\pm \mathbf{0.603} \times 10^5}$	$0.95374_{\pm 0.00815}$	$0.98058_{\pm 0.00400}$	$\mathbf{1.403}_{\pm \mathbf{0.702} \times 10^5}$	$0.95117_{\pm 0.00966}$
0.01	Bonferroni	$0.98999_{\pm 0.00269}$	$7.013_{\pm 4.265 \times 10^8}$	$0.97219_{\pm 0.00708}$	$0.99001_{\pm 0.00224}$	$3.086_{\pm 1.684 \times 10^8}$	$0.97184_{\pm 0.00602}$
	Mahalanobis	$0.99029_{\pm 0.00242}$	$8.583_{\pm 5.988 \times 10^6}$	$0.97326_{\pm 0.00625}$	$0.99037_{\pm 0.00231}$	$5.510_{\pm 3.444 \times 10^6}$	$0.97345_{\pm 0.00624}$
	Density (med/2)	$0.99013_{\pm 0.00242}$	$2.917_{\pm 0.845 \times 10^6}$	$0.97395_{\pm 0.00619}$	$0.99001_{\pm 0.00236}$	$1.727_{\pm 0.623 \times 10^6}$	$0.97254_{\pm 0.00606}$
	OT-CP	$0.99007_{\pm 0.00221}$	$5.059_{\pm 11.050 \times 10^8}$	$0.97360_{\pm 0.00529}$	$0.99016_{\pm 0.00272}$	$2.476_{\pm 5.087 \times 10^8}$	$0.97407_{\pm 0.00635}$
	MKS (med/2)	$0.98981_{\pm 0.00268}$	$\mathbf{5.538}_{\pm \mathbf{2.898} \times 10^5}$	$0.97324_{\pm 0.00613}$	$0.99000_{\pm 0.00241}$	$\mathbf{6.400}_{\pm \mathbf{3.344} \times 10^5}$	$0.97257_{\pm 0.00593}$

Table 12: House dataset ($d = 8$), Ridge and MLP. Mean $\pm$ std over 50 seeds.

House $d = 10$

		Ridge			MLP		
α	Method	Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	$0.90031_{\pm 0.00889}$	$3.655_{\pm 1.386 \times 10^{6}}$	$0.81452_{\pm 0.01774}$	$0.90040_{\pm 0.00893}$	$2.020_{\pm 0.835 \times 10^{6}}$	$0.81630_{\pm 0.01537}$
	Mahalanobis	$0.90081_{\pm 0.00825}$	$7.574_{\pm 2.510 \times 10^{4}}$	$0.79989_{\pm 0.01476}$	$0.90113_{\pm 0.00791}$	$4.971_{\pm 1.612 \times 10^{4}}$	$0.79086_{\pm 0.01549}$
	Density (med/2)	$0.90033_{\pm 0.00809}$	$1.936_{\pm 0.391 \times 10^{5}}$	$0.80688_{\pm 0.01583}$	$0.90084_{\pm 0.00781}$	$1.028_{\pm 0.238 \times 10^{5}}$	$0.79270_{\pm 0.01650}$
	OT-CP	$0.89941_{\pm 0.00773}$	$4.959_{\pm 1.200 \times 10^{5}}$	$0.80037_{\pm 0.01535}$	$0.90105_{\pm 0.00786}$	$3.446_{\pm 0.901 \times 10^{5}}$	$0.78951_{\pm 0.01602}$
	MKS (med/2)	$0.90120_{\pm 0.00741}$	$\mathbf{1.490}_{\pm \mathbf{0.312} \times 10^{4}}$	$0.79483_{\pm 0.01595}$	$0.90138_{\pm 0.00776}$	$\mathbf{1.687}_{\pm \mathbf{0.433} \times 10^{4}}$	$0.77516_{\pm 0.01757}$
0.05	Bonferroni	$0.95013_{\pm 0.00615}$	$5.299_{\pm 2.349 \times 10^{7}}$	$0.89619_{\pm 0.01215}$	$0.95058_{\pm 0.00596}$	$3.167_{\pm 1.450 \times 10^{7}}$	$0.89899_{\pm 0.01076}$
	Mahalanobis	$0.95020_{\pm 0.00548}$	$5.542_{\pm 1.925 \times 10^{5}}$	$0.88607_{\pm 0.01252}$	$0.95049_{\pm 0.00583}$	$4.396_{\pm 1.700 \times 10^{5}}$	$0.88487_{\pm 0.01288}$
	Density (med/2)	$0.95070_{\pm 0.00661}$	$2.092_{\pm 0.927 \times 10^{6}}$	$0.89584_{\pm 0.01338}$	$0.95075_{\pm 0.00613}$	$1.241_{\pm 0.527 \times 10^{6}}$	$0.89320_{\pm 0.01321}$
	OT-CP	$0.95075_{\pm 0.00570}$	$6.937_{\pm 3.723 \times 10^{6}}$	$0.89423_{\pm 0.01207}$	$0.95140_{\pm 0.00578}$	$5.535_{\pm 2.834 \times 10^{6}}$	$0.89109_{\pm 0.01271}$
	MKS (med/2)	$0.95031_{\pm 0.00611}$	$\mathbf{1.609}_{\pm \mathbf{0.597} \times 10^{5}}$	$0.89344_{\pm 0.01298}$	$0.95063_{\pm 0.00580}$	$\mathbf{1.914}_{\pm \mathbf{0.698} \times 10^{5}}$	$0.88564_{\pm 0.01335}$
0.02	Bonferroni	$0.98056_{\pm 0.00366}$	$5.367_{\pm 5.058 \times 10^{9}}$	$0.95215_{\pm 0.00807}$	$0.98058_{\pm 0.00374}$	$2.221_{\pm 1.959 \times 10^{9}}$	$0.95205_{\pm 0.00830}$
	Mahalanobis	$0.98051_{\pm 0.00345}$	$2.658_{\pm 1.630 \times 10^{7}}$	$0.95073_{\pm 0.00829}$	$0.98044_{\pm 0.00348}$	$1.552_{\pm 0.947 \times 10^{7}}$	$0.95047_{\pm 0.00837}$
	Density (med/2)	$0.98058_{\pm 0.00353}$	$5.692_{\pm 1.999 \times 10^{7}}$	$0.95207_{\pm 0.00817}$	$0.98086_{\pm 0.00366}$	$3.042_{\pm 1.084 \times 10^{7}}$	$0.95243_{\pm 0.00854}$
	OT-CP	$0.98078_{\pm 0.00330}$	$1.794_{\pm 2.425 \times 10^{9}}$	$0.95310_{\pm 0.00766}$	$0.98046_{\pm 0.00319}$	$1.108_{\pm 1.361 \times 10^{9}}$	$0.95111_{\pm 0.00749}$
	MKS (med/2)	$0.98051_{\pm 0.00370}$	$\mathbf{2.446}_{\pm \mathbf{1.144} \times 10^{6}}$	$0.95274_{\pm 0.00779}$	$0.98035_{\pm 0.00332}$	$\mathbf{2.590}_{\pm \mathbf{1.077} \times 10^{6}}$	$0.95088_{\pm 0.00819}$
0.01	Bonferroni	$0.98996_{\pm 0.00272}$	$1.052_{\pm 0.748 \times 10^{11}}$	$0.97249_{\pm 0.00711}$	$0.99028_{\pm 0.00228}$	$4.314_{\pm 2.690 \times 10^{10}}$	$0.97291_{\pm 0.00606}$
	Mahalanobis	$0.99022_{\pm 0.00258}$	$2.404_{\pm 2.124 \times 10^{8}}$	$0.97314_{\pm 0.00683}$	$0.99037_{\pm 0.00247}$	$1.554_{\pm 1.213 \times 10^{8}}$	$0.97380_{\pm 0.00660}$
	Density (med/2)	$0.99036_{\pm 0.00235}$	$1.618_{\pm 0.582 \times 10^{8}}$	$0.97414_{\pm 0.00598}$	$0.99034_{\pm 0.00249}$	$9.313_{\pm 3.698 \times 10^{7}}$	$0.97381_{\pm 0.00629}$
	OT-CP	$0.99031_{\pm 0.00225}$	$5.398_{\pm 14.428 \times 10^{10}}$	$0.97337_{\pm 0.00599}$	$0.99046_{\pm 0.00226}$	$3.249_{\pm 6.464 \times 10^{10}}$	$0.97409_{\pm 0.00593}$
	MKS (med/2)	$0.98980_{\pm 0.00276}$	$\mathbf{1.529}_{\pm \mathbf{0.979} \times 10^{7}}$	$0.97311_{\pm 0.00644}$	$0.99016_{\pm 0.00244}$	$\mathbf{1.708}_{\pm \mathbf{1.102} \times 10^{7}}$	$0.97287_{\pm 0.00596}$

Table 13: House dataset ($d = 10$), Ridge and MLP. Mean $\pm$ std over 50 seeds.

Blog $d = 2$

		Ridge			MLP		
α	Method	Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	$0.89928_{\pm 0.00431}$	$8.331_{\pm 0.869}$	$0.79801_{\pm 0.00845}$	$0.89926_{\pm 0.00413}$	$3.449_{\pm 0.980}$	$0.77705_{\pm 0.01435}$
	Mahalanobis	$0.89944_{\pm 0.00418}$	$6.752_{\pm 0.684}$	$0.79486_{\pm 0.00830}$	$0.89943_{\pm 0.00407}$	$2.894_{\pm 0.803}$	$0.77036_{\pm 0.01404}$
	Density (med/2)	$0.89934_{\pm 0.00416}$	$4.163_{\pm 0.110}$	$0.76854_{\pm 0.00866}$	$0.89914_{\pm 0.00456}$	$1.776_{\pm 0.514}$	$0.72959_{\pm 0.01327}$
	OT-CP	$0.89948_{\pm 0.00438}$	$4.537_{\pm 0.159}$	$0.78036_{\pm 0.00814}$	$0.89913_{\pm 0.00433}$	$1.852_{\pm 0.551}$	$0.73659_{\pm 0.01204}$
	MKS (med/2)	$0.90024_{\pm 0.00457}$	$\mathbf{3.049}_{\pm \mathbf{0.116}}$	$0.72292_{\pm 0.01147}$	$0.89905_{\pm 0.00491}$	$\mathbf{1.722}_{\pm \mathbf{0.432}}$	$0.72824_{\pm 0.01344}$
0.05	Bonferroni	$0.94954_{\pm 0.00301}$	$11.27_{\pm 1.08}$	$0.86834_{\pm 0.00760}$	$0.94967_{\pm 0.00270}$	$6.104_{\pm 1.450}$	$0.86324_{\pm 0.00943}$
	Mahalanobis	$0.94974_{\pm 0.00275}$	$9.182_{\pm 0.847}$	$0.86632_{\pm 0.00724}$	$0.94949_{\pm 0.00280}$	$5.121_{\pm 1.172}$	$0.86000_{\pm 0.00964}$
	Density (med/2)	$0.94986_{\pm 0.00363}$	$6.790_{\pm 0.324}$	$0.83967_{\pm 0.00922}$	$0.94961_{\pm 0.00306}$	$3.769_{\pm 0.784}$	$0.84140_{\pm 0.01016}$
	OT-CP	$0.94997_{\pm 0.00334}$	$6.975_{\pm 0.298}$	$0.84911_{\pm 0.00802}$	$0.94968_{\pm 0.00290}$	$3.859_{\pm 0.838}$	$0.84390_{\pm 0.00931}$
	MKS (med/2)	$0.94983_{\pm 0.00352}$	$\mathbf{5.664}_{\pm \mathbf{0.349}}$	$0.83123_{\pm 0.01056}$	$0.94943_{\pm 0.00287}$	$\mathbf{3.663}_{\pm \mathbf{0.747}}$	$0.84300_{\pm 0.00948}$
0.02	Bonferroni	$0.98001_{\pm 0.00191}$	$21.98_{\pm 2.51}$	$0.91833_{\pm 0.00692}$	$0.97960_{\pm 0.00209}$	$14.25_{\pm 3.13}$	$0.93366_{\pm 0.00797}$
	Mahalanobis	$0.98011_{\pm 0.00203}$	$19.43_{\pm 2.43}$	$0.91758_{\pm 0.00723}$	$0.97959_{\pm 0.00209}$	$11.98_{\pm 2.67}$	$0.93268_{\pm 0.00783}$
	Density (med/2)	$0.98011_{\pm 0.00213}$	$19.35_{\pm 1.61}$	$0.91607_{\pm 0.00813}$	$0.97978_{\pm 0.00213}$	$\mathbf{9.517}_{\pm \mathbf{1.822}}$	$0.92642_{\pm 0.00781}$
	OT-CP	$0.98006_{\pm 0.00208}$	$20.26_{\pm 2.44}$	$0.91673_{\pm 0.00768}$	$0.97984_{\pm 0.00201}$	$11.01_{\pm 1.99}$	$0.92591_{\pm 0.00736}$
	MKS (med/2)	$0.98024_{\pm 0.00216}$	$\mathbf{14.49}_{\pm \mathbf{1.20}}$	$0.91929_{\pm 0.00841}$	$0.97971_{\pm 0.00186}$	$9.762_{\pm 1.595}$	$0.92709_{\pm 0.00680}$
0.01	Bonferroni	$0.99023_{\pm 0.00166}$	$54.07_{\pm 9.15}$	$0.95837_{\pm 0.00700}$	$0.98978_{\pm 0.00155}$	$28.87_{\pm 9.79}$	$0.96221_{\pm 0.00538}$
	Mahalanobis	$0.99027_{\pm 0.00168}$	$44.74_{\pm 6.85}$	$0.95853_{\pm 0.00718}$	$0.98973_{\pm 0.00154}$	$23.79_{\pm 8.04}$	$0.96174_{\pm 0.00551}$
	Density (med/2)	$0.99038_{\pm 0.00148}$	$31.15_{\pm 2.73}$	$0.95870_{\pm 0.00591}$	$0.99000_{\pm 0.00143}$	$\mathbf{18.45}_{\pm \mathbf{3.31}}$	$0.96077_{\pm 0.00538}$
	OT-CP	$0.99034_{\pm 0.00165}$	$41.95_{\pm 4.65}$	$0.95846_{\pm 0.00708}$	$0.98987_{\pm 0.00145}$	$24.10_{\pm 4.42}$	$0.95860_{\pm 0.00544}$
	MKS (med/2)	$0.99018_{\pm 0.00152}$	$\mathbf{27.24}_{\pm \mathbf{2.96}}$	$0.95791_{\pm 0.00597}$	$0.98991_{\pm 0.00157}$	$20.23_{\pm 3.85}$	$0.96089_{\pm 0.00583}$

Table 14: Blog dataset ($d = 2$), Ridge and MLP. Mean $\pm$ std over 50 seeds.

Blog $d = 3$

		Ridge			MLP		
α	Method	Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	0.89990±0.00409	19.66±1.93	0.76751±0.00897	0.90011±0.00354	8.126±2.579	0.76696±0.00859
	Mahalanobis	0.90011±0.00479	13.75±1.27	0.75441±0.00935	0.90021±0.00399	6.115±2.416	0.76096±0.00828
	Density (med/2)	0.90003±0.00449	9.797±0.413	0.74079±0.00958	0.89965±0.00395	3.912±1.369	0.74125±0.00940
	OT-CP	0.90014±0.00465	11.01±0.52	0.75375±0.00804	0.89976±0.00407	4.148±1.538	0.74406±0.00921
	MKS (med/2)	0.89961±0.00454	**7.226±0.400**	0.71731±0.01019	0.89981±0.00454	**3.588±1.229**	0.73545±0.01101
0.05	Bonferroni	0.95017±0.00333	39.27±3.55	0.84628±0.00935	0.95035±0.00302	20.32±5.47	0.86750±0.00822
	Mahalanobis	0.95055±0.00337	28.26±2.57	0.84526±0.00912	0.95050±0.00320	14.67±5.02	0.86457±0.00922
	Density (med/2)	0.95028±0.00333	22.24±1.58	0.83698±0.00953	0.95021±0.00320	10.07±3.07	0.84942±0.00879
	OT-CP	0.95016±0.00357	22.27±1.53	0.84022±0.00941	0.95018±0.00295	10.66±2.92	0.85173±0.00880
	MKS (med/2)	0.95007±0.00336	**16.39±1.16**	0.83319±0.00960	0.95022±0.00314	**9.179±2.596**	0.84922±0.00903
0.02	Bonferroni	0.98025±0.00202	165.0±25.5	0.92998±0.00693	0.97987±0.00213	68.26±19.59	0.93681±0.00630
	Mahalanobis	0.98034±0.00214	103.3±14.6	0.92976±0.00752	0.97987±0.00207	45.47±16.65	0.93576±0.00682
	Density (med/2)	0.98044±0.00209	80.62±5.24	0.92581±0.00779	0.98008±0.00204	**28.90±8.02**	0.93006±0.00657
	OT-CP	0.98033±0.00241	97.59±15.24	0.92240±0.00823	0.98011±0.00219	37.62±9.10	0.93028±0.00679
	MKS (med/2)	0.98014±0.00233	**48.13±4.91**	0.92499±0.00793	0.98018±0.00201	31.63±7.28	0.93105±0.00670
0.01	Bonferroni	0.99027±0.00163	549.4±135.6	0.96450±0.00554	0.98998±0.00133	189.7±70.4	0.96491±0.00432
	Mahalanobis	0.99024±0.00166	312.3±75.7	0.96424±0.00563	0.98998±0.00144	119.9±60.0	0.96442±0.00478
	Density (med/2)	0.99018±0.00156	132.2±12.8	0.96105±0.00541	0.99000±0.00148	**61.51±15.51**	0.96197±0.00512
	OT-CP	0.99001±0.00168	236.9±40.3	0.95980±0.00674	0.98991±0.00152	106.0±31.5	0.96063±0.00500
	MKS (med/2)	0.99015±0.00158	**99.03±14.36**	0.95975±0.00594	0.99009±0.00166	69.76±14.12	0.96231±0.00599

Table 15: Blog dataset ($d = 3$), Ridge and MLP. Mean ± std over 50 seeds.

Blog $d = 4$

		Ridge			MLP		
α	Method	Coverage	Volume	WSC	Coverage	Volume	WSC
0.1	Bonferroni	0.89974±0.00401	34.62±11.46	0.76656±0.00844	0.89874±0.00387	13.80±8.96	0.76012±0.01132
	Mahalanobis	0.89960±0.00493	16.16±5.20	0.75140±0.00960	0.89900±0.00386	7.633±5.564	0.75081±0.01099
	Density (med/2)	0.89959±0.00453	12.27±0.74	0.74199±0.01030	0.89940±0.00417	4.406±3.027	0.73946±0.01176
	OT-CP	0.89979±0.00446	14.70±1.07	0.75743±0.00923	0.89965±0.00385	4.785±3.323	0.74234±0.01139
	MKS (med/2)	0.89926±0.00418	**3.472±0.394**	0.72208±0.01068	0.89952±0.00436	**2.975±2.373**	0.73056±0.01239
0.05	Bonferroni	0.95005±0.00326	93.53±31.98	0.84903±0.00828	0.94945±0.00368	49.45±27.95	0.86673±0.01019
	Mahalanobis	0.95025±0.00349	44.86±14.72	0.84720±0.00903	0.94933±0.00345	25.22±16.16	0.86221±0.00907
	Density (med/2)	0.95017±0.00342	39.85±4.15	0.84328±0.00946	0.94995±0.00312	14.96±9.15	0.85330±0.00834
	OT-CP	0.95002±0.00362	38.57±3.92	0.84628±0.00929	0.94974±0.00315	18.64±9.29	0.85553±0.00862
	MKS (med/2)	0.94993±0.00354	**11.67±1.66**	0.84322±0.00992	0.94978±0.00316	**10.35±6.60**	0.84911±0.00882
0.02	Bonferroni	0.97997±0.00224	744.6±330.3	0.93284±0.00709	0.97973±0.00228	306.3±166.2	0.93913±0.00651
	Mahalanobis	0.98017±0.00238	289.6±124.5	0.93326±0.00757	0.97985±0.00234	133.7±70.9	0.93774±0.00653
	Density (med/2)	0.98032±0.00214	201.3±15.7	0.92826±0.00793	0.97974±0.00275	**55.77±32.47**	0.93117±0.00851
	OT-CP	0.98021±0.00265	301.4±66.0	0.92573±0.00921	0.97968±0.00250	144.3±60.5	0.93098±0.00732
	MKS (med/2)	0.98017±0.00219	**50.72±8.09**	0.92998±0.00707	0.97980±0.00236	57.55±24.81	0.93213±0.00772
0.01	Bonferroni	0.99028±0.00167	$4.364_{\pm 1.954\times 10^{3}}$	0.96586±0.00559	0.98984±0.00183	$1.359_{\pm 0.825\times 10^{3}}$	0.96614±0.00512
	Mahalanobis	0.99033±0.00167	$1.491_{\pm 0.653\times 10^{3}}$	0.96589±0.00562	0.99005±0.00172	558.5±342.9	0.96602±0.00522
	Density (med/2)	0.99005±0.00177	363.4±44.6	0.96173±0.00629	0.98980±0.00171	**159.7±69.7**	0.96334±0.00573
	OT-CP	0.98999±0.00188	$1.680_{\pm 0.859\times 10^{3}}$	0.96151±0.00694	0.99008±0.00190	978.5±522.0	0.96237±0.00612
	MKS (med/2)	0.99001±0.00176	**153.9±40.2**	0.96228±0.00598	0.98967±0.00183	171.3±61.7	0.96307±0.00582

Table 16: Blog dataset ($d = 4$), Ridge and MLP. Mean ± std over 50 seeds.

C.4.4 Hyperparameter ablation

We assess sensitivity to both hyperparameters jointly, sweeping the lengthscale ℓ over nine values spanning a $256\times$ range and the regularization γ over seven percentiles of the centred Gram spectrum, on three datasets. Coverage is unaffected throughout, which is the empirical counterpart of the guarantee in Theorem 3.4, valid for any configuration. Volumes vary with both parameters but are stable across a wide interior region and the only configurations that lose to the Mahalanobis ellipsoid combine a very wide kernel with a very small γ. A wide kernel collapses the centred spectrum onto

a few smooth directions, and a low-percentile γ then lands among the near-zero eigenvalues below them, so the score approaches the unregularized limit in which the calibration scores γh_{ii} degenerate (Proposition 4.6).

	Volume ratio to Mahalanobis						
ℓ / (med/2)	5	10	30	50	70	90	95
16	4.139	1.953	0.665	0.700	0.782	0.904	0.891
8	1.537	0.907	0.594	0.623	0.669	0.773	0.819
4	0.648	0.530	0.533	0.554	0.587	0.652	0.683
2	0.490	0.490	0.492	0.503	0.522	0.562	0.585
1	0.499	0.498	0.490	0.487	0.487	0.505	0.518
1/2	0.521	0.522	0.517	0.511	0.513	0.496	0.501
1/4	0.629	0.629	0.636	0.641	0.642	0.637	0.632
1/8	0.702	0.702	0.701	0.701	0.700	0.699	0.698
1/16	0.712	0.712	0.712	0.712	0.712	0.712	0.712

	Coverage						
ℓ / (med/2)	5	10	30	50	70	90	95
16	0.9016	0.8994	0.9005	0.9008	0.9005	0.9001	0.9006
8	0.9009	0.8990	0.9013	0.9012	0.9008	0.9005	0.9004
4	0.9017	0.9021	0.9020	0.9016	0.9013	0.9010	0.9010
2	0.9019	0.9018	0.9018	0.9022	0.9024	0.9020	0.9021
1	0.9019	0.9018	0.9023	0.9022	0.9026	0.9025	0.9021
1/2	0.9021	0.9023	0.9024	0.9023	0.9020	0.9021	0.9023
1/4	0.9020	0.9019	0.9021	0.9022	0.9023	0.9023	0.9024
1/8	0.9023	0.9023	0.9023	0.9023	0.9023	0.9022	0.9022
1/16	0.9024	0.9024	0.9024	0.9024	0.9024	0.9024	0.9024

Table 17: House d=4, $1-\alpha$=0.9, 30 seeds. Rows: kernel lengthscale as a multiple of the median heuristic $\mathrm{med}/2$. Columns: γ at the given percentile of the centred-Gram spectrum. Left: volume ratio to Mahalanobis. Right: empirical coverage, target 0.9, which is unaffected by either parameter.

	Volume ratio to Mahalanobis						
ℓ / (med/2)	5	10	30	50	70	90	95
16	4.078	1.548	0.570	0.610	0.658	0.741	0.818
8	2.349	0.948	0.521	0.537	0.572	0.644	0.687
4	1.052	0.551	0.484	0.492	0.510	0.535	0.558
2	0.635	0.458	0.459	0.463	0.470	0.482	0.494
1	0.469	0.447	0.446	0.446	0.449	0.453	0.458
1/2	0.448	0.447	0.444	0.442	0.442	0.443	0.444
1/4	0.442	0.443	0.452	0.449	0.445	0.443	0.441
1/8	0.440	0.441	0.453	0.473	0.462	0.455	0.451
1/16	0.520	0.521	0.526	0.545	0.551	0.533	0.522

	Coverage						
ℓ / (med/2)	5	10	30	50	70	90	95
16	0.9008	0.8998	0.8990	0.8990	0.8990	0.8989	0.8990
8	0.8998	0.8987	0.8991	0.8987	0.8988	0.8988	0.8987
4	0.8989	0.8994	0.8994	0.8989	0.8990	0.8985	0.8986
2	0.8993	0.8995	0.8997	0.8996	0.8996	0.8991	0.8990
1	0.9000	0.9001	0.8998	0.8998	0.9000	0.8999	0.8998
1/2	0.9001	0.9002	0.9001	0.9001	0.9003	0.9001	0.9000
1/4	0.9000	0.8999	0.9001	0.9002	0.9001	0.9001	0.9000
1/8	0.8999	0.8997	0.8999	0.9001	0.8997	0.9000	0.8998
1/16	0.8993	0.8993	0.8994	0.8996	0.8996	0.8995	0.8995

Table 18: Blog d=2, $1-\alpha$=0.9, 30 seeds. Rows: kernel lengthscale as a multiple of the median heuristic $\mathrm{med}/2$. Columns: γ at the given percentile of the centred-Gram spectrum. Left: volume ratio to Mahalanobis. Right: empirical coverage, target 0.9, which is unaffected by either parameter.

	Volume ratio to Mahalanobis						
ℓ / (med/2)	5	10	30	50	70	90	95
16	2.805	2.254	0.559	0.573	0.601	0.645	0.745
8	1.556	0.963	0.531	0.543	0.566	0.626	0.641
4	0.937	0.652	0.503	0.513	0.529	0.566	0.590
2	0.604	0.498	0.483	0.489	0.496	0.516	0.532
1	0.463	0.461	0.462	0.464	0.474	0.483	0.490
1/2	0.450	0.449	0.446	0.448	0.447	0.451	0.458
1/4	0.450	0.449	0.450	0.453	0.459	0.442	0.440
1/8	0.518	0.518	0.521	0.528	0.527	0.521	0.518
1/16	0.559	0.559	0.559	0.559	0.559	0.559	0.559

	Coverage						
ℓ / (med/2)	5	10	30	50	70	90	95
16	0.8998	0.8992	0.8992	0.8995	0.8995	0.8992	0.8996
8	0.9006	0.9000	0.8991	0.8993	0.8993	0.8992	0.8990
4	0.8989	0.8992	0.8988	0.8988	0.8989	0.8990	0.8992
2	0.8995	0.8983	0.8989	0.8990	0.8990	0.8989	0.8986
1	0.8984	0.8984	0.8983	0.8983	0.8988	0.8989	0.8988
1/2	0.8989	0.8987	0.8987	0.8987	0.8985	0.8988	0.8988
1/4	0.8983	0.8984	0.8989	0.8990	0.8989	0.8988	0.8987
1/8	0.8997	0.8998	0.8999	0.8998	0.8998	0.8998	0.8997
1/16	0.8999	0.8999	0.9000	0.9000	0.9001	0.9000	0.9000

Table 19: Bio d=3, $1-\alpha$=0.9, 30 seeds. Rows: kernel lengthscale as a multiple of the median heuristic $\mathrm{med}/2$. Columns: γ at the given percentile of the centred-Gram spectrum. Left: volume ratio to Mahalanobis. Right: empirical coverage, target 0.9, which is unaffected by either parameter.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three contributions: the kernel score definition with GP identity (Propositions 3.3 and 4.1), coverage guarantees with effective rank convergence rates (Theorem 3.4), and empirical volume reduction (Section 5). All are supported by theoretical results and experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 2 states that the regions have a fixed shape and do not adapt to the covariates. Section 5 defers the choice and tuning of the kernel to future work, and notes that Monte Carlo volume estimation limits the experiments to output dimensions up to $d = 10$. The theoretical guarantees require exchangeability or α-mixing (Section 3, Appendix B).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results are numbered and cross-referenced. Assumptions are stated in Appendix B. Full proofs for all propositions and theorems are provided in Appendices A and B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The data generating process, model architectures, data splits, hyperparameter heuristics, number of seeds, and evaluation metrics are fully specified in Section 5 and Appendix C. The code repository is provided for reproducibility. All datasets are publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets are publicly available and cited. The code and the raw results behind every table are available at `https://github.com/meyerlouis/mks-conformal`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 and Appendix C specify all details including data splits, model architectures, hyperparameter selection heuristics, kernel choice, lengthscale values, and evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All main results report mean $\pm$ standard deviation across 30 random seeds (synthetic) and 50 random seeds (real data). The source of variability is the random data generation and split (synthetic) and the random split (real data).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Experiments were run on CPUs. The method itself is not computationally intensive; multiple cores were used for computing evaluation metrics across random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is theoretical and methodological, uses only publicly available datasets, and involves no human subjects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is foundational statistical methodology for uncertainty quantification with no direct path to negative societal applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: No models or datasets with risk of misuse are released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The three datasets (House, Bio, Blog) are publicly available and cited. The OT-CP baseline is a port of OTT-JAX, released under the Apache 2.0 license, as stated in the code repository.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets or models are released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects research was conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: No human subjects research was conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as a component of the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.